SUPPORTING HUMAN INTERPRETATION AND ANALYSIS OF ACTIVITY CAPTURED THROUGH OVERHEAD VIDEO

A Dissertation
Presented to
The Academic Faculty

by

Mario Romero

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Computer Science

Georgia Institute of Technology

August 2009

SUPPORTING HUMAN INTERPRETATION AND ANALYSIS OF ACTIVITY CAPTURED THROUGH OVERHEAD VIDEO

Approved by:

Dr. Gregory Abowd, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. John Stasko
School of Interactive Computing
Georgia Institute of Technology

Dr. Elle Yi-Luen Do
School of Interactive Computing and
College of Architecture
Georgia Institute of Technology

Dr. John Peponis
College of Architecture
Georgia Institute of Technology

Dr. James Foley
School of Interactive Computing
Georgia Institute of Technology

Date Approved: June 24, 2009

Para mi Nata y mi Gabo.

ACKNOWLEDGEMENTS

This document is a testament of support. My greatest support and love is my family, Natalia and Gabriel. They are my infinite source of inspiration and motivation. Natalia shares this life with me with absolute devotion and joy. Gabriel's sparks of pure happiness light up the next step at every step.

My advisor, Dr. Gregory Abowd, is the modest genius who brings out the very best in everybody. I am a better person by his guidance and example.

I have a deep admiration for the personal and professional accomplishments of my thesis committee, Dr. Ellen Do, Dr. Jim Foley, Dr. John Stasko, and Dr. John Peponis. Their advice sets a standard that I will continue to pursue. I would like to specially thank the honorary member of my thesis committee, Dr. Rosa Arriaga, who has shared her best at every step of this work.

My friends and colleges at the Ubiquitous Computing Laboratory have created a superb thinking and supportive environment. I wish that at every stage of the laboratory its members will benefit from the same stimulating friendship I have received. I wish to thank Erich Stutenbeck, Lana Yarosh, Adebola Osuntogun, Tae-Jung Yun, Hee Young Jeong, Fatima Boujarwah, Nazneen Anwer, Shwetak Patel, Julie Kientz, Gillian Hayes, Aras Bilgen, and Giovanni Iachello for their friendship and motivation. I would like to specially thank Jay Summet, Tracy Westeyn, Kris Nagel, Dounia Berrada, Sergio Goldenberg, Alice Vialard, Mukil Kesavan, and Kate Rosier. I extend my warmest gratitude to Dr. Maureen Biggers who gave me her open hand when I most needed it.

I extend my vow of gratitude to the people who make it work at GVU, the School of Interactive Computing, and the Aware Home Research Initiative. I must thank profusely all the participants of my studies as well.

Finally, this work is the result of lifelong learning. The foundations for this and all achievements exist only because of my mother, Guiomar Vega. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

Page

Page

Page

Page

Page

Page

LIST OF ABBREVIATIONS

3D-FP	Three-Dimensional Floor Plan
AC	Activity Cube
AI	Artificial Intelligence
AM	Activity Map
AT	Activity Table
AFD	Adjacent Frame Difference
AMP	Active Multiple Projection
DFP	Digital Family Portrait
EDF	Social Energy, Density, and Flow
FPS	Frames per Second
GIS	Geographical Information System
HCI	Human-Computer Interaction
OHC	Overhead Camera
OHV	Overhead Video
PMP	Passive Multiple Projection
RFID	Radio Frequency Identification
SAZ	Semantic Activity Zone
SP	Single Projection
TM	Tableau Machine
Viz-A-Vis	Visualization of Activity through Vision
VRP	Virtual Rear Projection
VC	Video Cube
VP	Video Playback

SUMMARY

Many disciplines spend considerable resources studying behavior. Tools range from pen-and-paper observation to biometric sensing. A tool's appropriateness depends on the goal and justification of the study, the observable context and feature set of target behaviors, the observers' resources, and the subjects' tolerance to intrusiveness. We present two systems: Viz-A-Vis and Tableau Machine. Viz-A-Vis is an analytical tool appropriate for onsite, continuous, wide-coverage and long-term capture, and for objective, contextual, and detailed analysis of the physical actions of subjects who consent to overhead video observation. Tableau Machine is a creative artifact for the home. It is a long-lasting, continuous, interactive, and abstract Art installation that captures overhead video and visualizes activity to open opportunities for creative interpretation.

We focus on overhead video observation because it affords a near one-to-one correspondence between pixels and floor plan locations, naturally framing the activity in its spatial context. Viz-A-Vis, or **VI**suali**Z**ation of **A**ctivity through **VIS**ion, is an information visualization interface that renders and manipulates computer vision abstractions. It visualizes the hidden structure of behavior in its spatiotemporal context. We demonstrate the practicality of this approach through two user studies. In the first user study, we show an important search performance boost when compared against standard video playback and against the video cube, an advanced technique from the video visualization literature. Furthermore, we determine a unanimous user choice for overviewing and searching with Viz-A-Vis. In the second study, a domain expert evaluation, we validate a number of real discoveries of insightful environmental behavior

patterns by a group of senior architects using Viz-A-Vis. Furthermore, we determine clear influences of Viz-A-Vis over the resulting architectural designs in the study.

Tableau Machine is a sensing, interpreting, and painting artificial intelligence. It is an Art installation with a model of perception and personality that continuously and enduringly engages its co-occupants in the home, creating an aura of presence. It perceives the environment through overhead cameras, interprets its perceptions with computational models of behavior, maps its interpretations to generative abstract visual compositions, and renders its compositions through paintings on a television screen and a printer. We validate the goal of opening a space for creative interpretation, playful experimentation, reflection, contemplation, and conversation through a study that included three long-term deployments in real family homes.

CHAPTER 1

INTRODUCTION AND MOTIVATION

Human activity is the subject of systematic study for a number of fields with varying goals. In computer science, researchers study the factors that afford or constrain physical and procedural interaction between humans and computer systems (Card, Moran et al. 1983; Pantic, Pentland et al. 2007). In architecture, researchers focus on the relationships between the environment and people's behavior (Proshansky 1976; Whyte 1980; Bechtel 1997). In behavioral therapy, practitioners track the internal and external causes and effects of target behaviors (Grant and Evans 1994). In security, operators filter normal actions striving to detect patterns of outlier behavior posing potential threats (Snidaro, Micheloni et al. 2005). In vehicular design, human-factors engineers concentrate on the patterns of effective, efficient, and affective appropriation of the design features of the cabin (Quan, Dong-chi et al. 2005). In manufacturing, industrial engineers measure work motion and time to increase productivity, safety, and quality (Barnes 1980). In retail, store managers study shoppers' behaviors to maximize space marketability (Underhill 2000). In video game creation, character designers study human activity to create behavioral models for their AI agents (Champandard 2003). Despite ubiquitous interest in human activity, current data collection practices lack substantial context, detail, fidelity, automation, continuity, and duration.

The methods of observation range widely, from non-intrusive, low-fidelity, episodic, subjective observation to intrusive, high fidelity, continuous, and objective sensing. The tools range from pen-and-paper observation (Barley 1990) to biometric

sensing (Prabhakar, Pankanti et al. 2003). The appropriateness of a tool depends on the analytical goals of the study, the observable features of target activities, the perceived benefits, length, and place of the study, the observer's resources, and the subjects' tolerance to intrusiveness.

The first overarching goal of this research is to provide activity analysts with practical mechanisms to interpret and analyze detailed activity continuously occurring widely across natural settings over variable periods, from a few seconds to many months. We have designed, developed, and evaluated Viz-A-Vis, or **VI**suali**Z**ation of **A**ctivity through **VIS**ion, in order to accomplish and validate this first goal. The second overarching goal of this research is to create a space for creative interpretation of human activity. We have designed, developed, longitudinally deployed, and evaluated Tableau Machine, a sentient Art installation, in order to accomplish and validate this second goal.

We define *activity interpretation* as the process of understanding sufficient detail and context directly from abstract aggregations, thus saving considerable time when compared to raw video browsing. We define activity analysis as the methodical process of finding target behaviors and events, measuring observable features in target behaviors, and synthesizing categories of behavioral patterns in order to discover methodically the relationship between stimuli, behavior, and consequences.

We define *creative interpretation* as the negotiated meaning-making process between the creator of an artifact, the emitter of the message, and the consumer of the artifact, the receiver of the message. Creative interpretation is the central theme of contemporary Art. By opening a space for creative interpretation, we also broaden the

scope of understanding and representing human activity. We open a possibility for creative discovery.

Concretely, we deploy overhead camera (OHC) networks and employ computer vision and information visualization techniques to abstract and visualize relevant video features. OHCs are appropriate for long-term and natural-environment, discrete, continuous, and detailed analysis of physical activities of subjects who tolerate overhead cameras. Through abstraction, it is also conductive of continuous and unobtrusive creative interpretation.

We utilize overhead video (OHV) as an event-capturing tool that has the coverage and contextual mappings to open new perspectives into everyday human behavior. Today's video collection methods typically place a number of front-view or corner-view cameras that do little to contextualize the wide spatial range of the activity they capture. Furthermore, the computer vision techniques that easily highlight relevancy in overhead video do not apply to other types of video. Other types of video require increasingly complex vision methods. Overhead cameras are the highest-resolution, widest-coverage, activity-capturing sensors that are practically deployable in natural environments over long periods and from which relevant aggregates can be easily computed. Because of their configuration, they readily afford robust low-level computer vision and, more importantly, pixel-level localization. OHV supports numerous potential applications. It can afford the long-term objective evaluation of behavioral therapy in special classrooms. It can track infant gross-motor developmental progress in the nursery. OHV can provide objective long-term and continuous measures of patients' movements under physical therapy in their natural environments, not just in the doctor's office. It can quantify

minutely and continuously long-term occupancy for the analysis and evaluation of architectural and vehicular interior designs. OHV can trace factory operations to increase industrial productivity and safety. It can discover customer behaviors to increase retail space marketability and security. It can track players and team motions across a game or a season to improve strategy and performance, illustrate the nuances of a play to a broadcast audience, and gather irreplaceable data to automatically build rich behavioral models for AI players and coaches in a video game version of the sport.

The world is already widely covered by overhead video, from satellites to security cameras to microscopes. Yet, today, these cameras do not automatically build spatiotemporal models of activity. They still rely heavily on tedious, onerous, and error-prone human-operator inspection. Overhead video or computed inferences from OHV have not crystallized these opportunities. The main reason is that despite its great potential, continuously recording video introduces important challenges. First, it rapidly generates overwhelmingly large data sets for manual inspection. Second, automatic and reliable high-level activity analysis in unconstrained natural places and periods remains an impractical goal for computer vision. There remains a large semantic gap between collections of video and analytic or interpretative insight. In other words, there is a representational difference between volumes of video and human-level understanding of the events in the video. We will discuss the semantic gap in chapter 4. Third, video intrudes on privacy.

In this thesis, we explicitly address the first two challenges. For the third challenge, we have implemented a number of privacy preserving techniques, such as the immediate deletion of original video frames, explicit outlining of the cameras' field of

view, user control to stop data collection or retroactively delete data, and the creation of a physical blur filter (Neustaedter, Greenberg et al. 2005; Hayes 2006). Nevertheless, Iachello's principle of Proportionality (Iachello and Abowd 2005) provides the most satisfactory answer to the privacy concerns afflicting all sensing technologies. Simply stated, to allow sensor intrusion is the subject's decision. The decision compares the weight of the perceived benefits versus the weight of the potential risks. Overhead cameras may be ethically employed only for those subjects, applications, places, and periods where the perceived benefits outweigh the potential risks due to loss of privacy.

Currently, only extensive playback bridges the semantic gap between large video collections and insight. The focus of this work is to efficiently bridge the semantic gap between low-level sensor data and high-level insight through a mixed-initiative computing approach (Allen, Guinn et al. 1999). From the machine's side, we bridge the gap with simple computer vision techniques that robustly aggregate and highlight relevant features in the raw data. From the human's side, we bridge the gap with interactive information visualization techniques that contextualize the relevant features while providing direct indexing to the raw video.

We simplify the computer vision and provide raw video indexing to sustain *accountability* (Button and Dourish 1996), *reification* (Gunderson and Gunderson 2008), and *learnability* (Valiant 1984; ISO 2001). When the machine fails to deliver correct or sufficient results, the human must be able to readily recognize and understand the failure in order to recover from it. The system needs to be accountable. We support recovery by providing indexed access to original video frames. That is reification, the reverse process of abstraction. Finally, humans are good pattern recognizers. They see patterns in clouds

and stars. By keeping the abstraction-reification process simple, users can easily learn to recognize activity in the machine vision abstractions.

1.1. Purpose of Research and Thesis Statement

In order to contextualize our thesis statement, we will briefly introduce a number of definitions here. We will revisit these definitions in detail in chapters 5 and 6.

The general goal of the Viz-A-Vis research is to create practical tools that support human interpretation and analysis of enduring activity in natural settings. This work focuses on two types of analytical tasks. First, it centers on descriptive tasks, where the goal is to gather direct evidence for a second-stage inductive analysis (Grant and Evans 1994). Second, it concentrates on behavior pattern discovery, a high-level analytical task that includes descriptive tasks and second-level analytical tasks, such as classifying, grouping, and synthesizing (Lofland and Lofland 1995).

In behavioral analysis, quantifying the *frequency*, *duration*, *latency*, and *percentage correct* are the primary descriptive tasks when focusing on behavior modification (Grant and Evans 1994). Since we are not focused on behavior modification, we have translated the four tasks above to these five tasks: (1) *describing;* (2) *bounding;* (3) *searching;* (4) *counting;* and (5) *tracking.* We based this translation on Grant's description and on the low-level components of analytic activity in information visualization described in (Amar, Eagan et al. 2005). We will discuss further the choice of these tasks in chapter 4.

Briefly, *describing* is the task of observing and verbalizing relevant features of activity captured in video. *Bounding* is the task of finding the time of start and end of an activity, regardless of its location. *Searching* is the task of locating in space and time

instances of specific target actions, behaviors, or events. *Counting* is the task of enumerating the repetitions of a target action. *Tracking* is following the location and describing the actions of a target subject. It is a description refined to include only a single subject across space and time.

For the tasks above, we apply the following human operator performance metrics: (1) *time to task completion;* (2) *precision;* (3) *recall;* and (4) *coverage*. *Time to task completion* is the period between the start and end of a task, including its subtasks, and the self-evaluation of results until the operator is satisfied. *Precision* is the percentage of correct instances from the set of retrieved instances. *Recall* is the percentage of retrieved instances from the set of target instances in the original video. *Coverage* is the length of video traversed during the task. We apply the metrics from the general information retrieval literature (Manning and Schütze 2002).

We define behavior pattern discovery as the task of systematically gathering and classifying evidence in the support of a theory connecting the causes, effects, and observable features of newly witnessed behaviors.

Using this language, our goals with Viz-A-Vis are: (1) to increase searching and bounding precision, recall, and coverage, thus lowering time to task completion; (2) to increase the view of activity across time in order to provide new overview vocabulary for the description and tracking of activity; (3) to provide a visual dictionary of behavior patterns across everyday episodes in order to facilitate new behavior pattern discovery; and (4) to improve the user experience by providing new perspectives of everyday life. Here the user is the activity analyst.

The second overarching goal of this thesis is to open opportunities for creative interpretation of activity. With Tableau Machine, we present home activity in new and unexpected ways to the occupants of the home. Tableau Machine is an Art installation that senses behavior using overhead cameras, aggregates and classifies the sensor data, and maps it to an artistic visual composition generator. Physically, it is a set of overhead cameras, a computer, a printer, and a large display. Our goal with Tableau Machine is to open opportunities for creative interpretation, playful experimentation, conversation, contemplation, and reflection regarding everyday life, rhythms, and activities. Our aim is to facilitate a space for reflective conversation about activity in people's homes.

To accomplish these goals, we propose the following *thesis statement*:

> *In the process of overhead video interpretation and analysis of activity, combining computer vision abstractions with information visualization techniques provides: (1) improved user task performance measured by time to task completion, precision, recall, coverage, and user assessment; (2) improved user experience measured by user preference; (3) increased user capacity to discover activity patterns; and (4) new opportunities for creative interpretation, experimentation, conversation, and reflection regarding everyday activities.*

To test this thesis statement, we built and evaluated Viz-A-Vis, a visualization of activity through computer vision, and Tableau Machine, a perceptive Art installation. Viz-A-Vis tests the first three claims. Tableau Machine tests the fourth and last claim. We ran five user studies, three for Viz-A-Vis, one formative and two summative, and two for Tableau Machine, one formative and one summative. The formative user study for Viz-A-Vis tested three prototypes through expert evaluation. The final version of Viz-A-

Vis emerged from this study. We describe its details in chapter 4. The first summative study with Viz-A-Vis evaluated activity analysts' task-based performance and preference. We compared Viz-A-Vis against traditional video playback and against the video cube, a sample of the state-of-the-art proposed in the video visualization literature (Fels, Lee et al. 2000; Klein, Sloan et al. 2002; Daniel and Chen 2003). We detail this study in chapter 5. The second summative study with Viz-A-Vis, a domain expert study, assessed its ability to open opportunities to methodically discover activity patterns among a group of architects. We discuss this study in chapter 6. Finally, we designed the final version of Tableau Machine through a formative study at Georgia Tech's Aware Home and we ran three in-home longitudinal studies. We report the results of these studies in chapter 3.

1.2. Research Questions

With this thesis, we address the following broad research questions:

- Can computer vision abstractions and information visualization techniques improve the interface to analyzing activity in overhead video as measured by time to task completion, precision, recall, coverage, and user assessment (Thesis Claim 1)?
- Can computer vision abstractions and information visualization techniques improve the user experience of activity video-analysis as measured by user preference (Thesis Claim 2)?
- Can vision-based data abstractions improve the information visualization interface as measured by analytical discovery of activity patterns (Thesis Claim 3)?

- Can a vision-based visualizing Art installation engage users in a long-term process of creative interpretation, experimentation, conversation, and reflection (Thesis Claim 4)?

Table 1.1 summarizes the four research questions above, the validation methods, and the location for each in this document.

Table 1.1: Summary of the thesis claims validations

Research Question	Validation
Can computer vision abstractions and information visualization techniques improve the interface to analyzing activity in overhead video as measured by time to task completion, precision, recall, coverage, and user assessment (Thesis Claim 1)?	Empirical usability test measuring user performance with Viz-A-Vis compared against (A) video playback and (B) the video cube (Section 5.3)
Can computer vision abstractions and information visualization techniques improve the user experience of activity video-analysis as measured by user preference (Thesis Claim 2)?	Empirical usability test measuring user preference with Viz-A-Vis compared against (A) video playback and (B) the video cube (Section 5.3)
Can vision-based data abstractions improve the information visualization interface as measured by analytical discovery of activity patterns (Thesis Claim 3)?	Empirical usability test and focus group with domain experts (Section 6.4).
Can a vision-based visualizing Art installation engage users in a long-term process of creative interpretation, experimentation, conversation, and reflection (Thesis Claim 4)?	Long-term, in-situ user study (Sections 3.1 and 3.2).

1.3. Thesis Overview

We present the development and evaluation of two perception-visualization systems, Tableau Machine, in chapter 3, and Viz-A-Vis, in chapter 4. Tableau Machine served as the motivation for the creation of Viz-A-Vis. We built the first versions of Viz-A-Vis during the design of Tableau Machine in order to find patterns of activity in the home.

Chapter 5 describes the design, analysis, and results of the user study matching Viz-A-Vis against two conditions: (A) traditional video playback and (B) the video cube, an advanced 3D visualization of video. Chapter 6 describes the study of discovery potential of Viz-A-Vis in the hands of a group of architects. Chapter 7 presents the conclusions of this thesis and the possible future work for this research.

Appendix A presents a formal mathematical definition of proxies for Social Energy, Density, and Flow. Finally, appendix B presents a glossary for lexica introduced in this thesis or ambiguous terms borrowed from related work.

CHAPTER 2

BACKGROUND AND RELATED WORK

Both Tableau Machine and Viz-A-Vis are multi-disciplinary systems. They build on concepts, theories, and methodologies from Artificial Intelligence, Human-Computer Interaction, Art, Information Visualization, Computer Vision, and Ubiquitous Computing. This chapter presents, field by field, the foundations and the related work for both systems.

2.1. The Semantic Gap and Mixed-Initiative Computing

The *semantic gap* is the difference between two representations with varying degree of abstraction of the same concept (Shih 2002; Hare, Lewis et al. 2006). In video, the representation of concepts is a sequence of digital images. Each image is a quantization of incoming light stored as a number matrix devoid of intrinsic meaning. Humans can readily abstract meaning from raw images into high-level semantics, such as natural language descriptions.

Alas, the automation of image understanding is a complex and brittle busyness. Furthermore, human understanding of video requires animation and that takes time. Long video sequences render manual inspection prohibitively lengthy. Thus, a semantic gap exists between voluminous image sequences and human understanding that currently can be bridged only through vast and monotonous observation. Moreover, given the extremely sparse distribution of target events that continuous video sequence presents, most of the invested time does not return rewards.

Humans use natural language, structured symbolic representations, to concisely describe and understand phenomena. In their natural use, these representations are very abstract, that is, they efficiently encode relevant information and eliminate unnecessary detail. Natural language representations are semantically packed. On the other hand, computational representations of raw images are burdened by unnecessary detail. Raw pixels most often include vast and irrelevant detail. In fact, by far, the most common scenario is for most of the pixels to be irrelevant to high-level analytical tasks. For instance, we estimate that in the nearly 13,000 hours of continuous overhead video we have collected in two offices, two living laboratories, four real homes, and two museums, roughly 1 out of every 1000 bits contains activity information, target or otherwise. In other words, 99.9% of the data is static background. Another example comes from behavioral therapists who use video for tracking patients' development (Abowd, McGee et al. 2009). The therapists do not run continuous video. Rather, they only sample a few hours throughout the days and weeks of therapy. Nevertheless, practitioners recount spending most of their time discarding true negatives. In other words, they spend most of their time searching for sparse and unpredictable target events.

Simply stated, computer vision and information visualization share the goal of segregating meaningful information entangled in raw data. The two fields differ because of the mechanisms used to achieve this goal, which are reflections of the central assumption of where reasoning occurs. In computer vision, the reasoning occurs in the machine through statistical pattern finding and reasoning algorithms. In information visualization, the reasoning occurs in the human analyst's cognitive and perceptual structures, and is augmented through interactive visualization, navigation, and

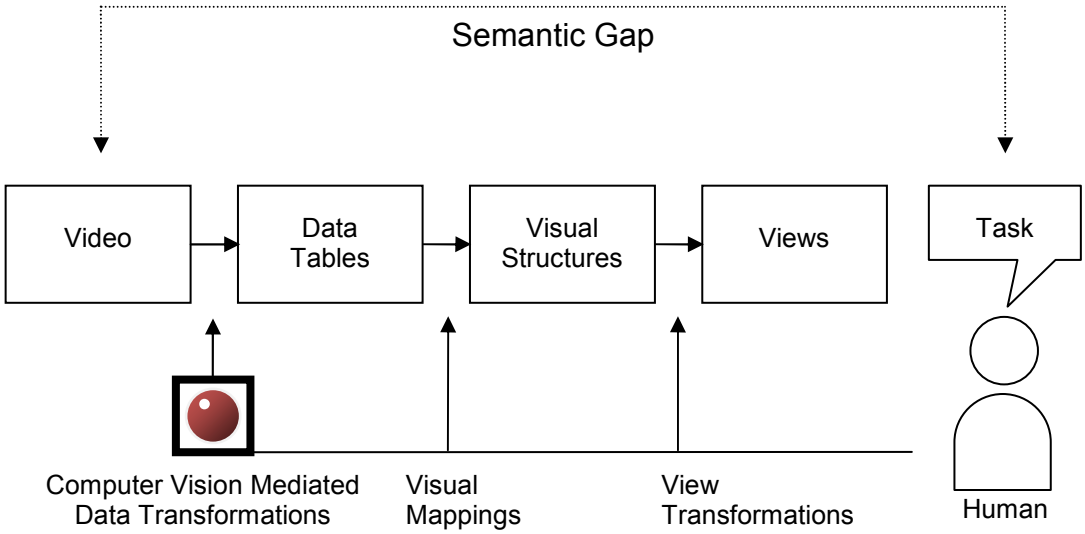


Figure 2.1: Traditional information visualization procedural model (Card, Mackinlay et al. 1999) augmented with automatic low-level data transformations from computer vision. High-level analysis remains in the human.

manipulation tools. An important theme of this thesis is exploring the rich potential for collaboration between the two fields. We use low-level computer vision to hide most of the unnecessary detail in the raw data, but purposely avoid higher-level abstractions that introduce complexity and brittleness into the process. As illustrated in Figure 2.1, our model of information visualization keeps the human at the core of reasoning and places the computer vision and pattern recognition at the data transformation level.

Similar to (Hare, Lewis et al. 2006), we share the goal of bridging the semantic gap between raw images and insight through mixed-initiative interaction (Allen, Guinn et al. 1999). From the human's side, we employ established information visualization techniques. From the computer's side, we perform automatic data transformations using computer vision. Our raw data is video. Tabularizing video without abstraction is equivalent to representing each pixel in a frame as an independent variable across time. For modest size frames, this representation is a time series with several hundred thousand

variables that is prohibitively expensive to store, process, and analyze. In practice, each pixel in an image sequence is not an independent variable. Pixels share high luminance and chrominance correlation with their spatial and temporal neighbors. Furthermore, the vast majority of pixels in an image stream from a static camera are irrelevant because they are identical to pixels in previous frames of the sequence, their temporal neighbors. We take advantage of these inherent properties of overhead video to automatically compute two robust low-level vision abstractions: motion (Adelson and Bergen 1985) and aggregation (Sonka, Hlavec et al. 1999). The trick is that they do not always share semantic correlation with their neighbors, so our spatial and temporal aggregates include human input to define more clearly semantic, rather than simply chrominance and luminance, boundaries. In chapters 3 and 4 we describe in detail the abstraction procedures underlying Tableau Machine and Viz-A-Vis.

2.2. Context-Aware Computing and Activity Recognition

Tableau Machine is an instance of a context-aware computing system (William, Robert et al. 2003), where the contextual sensing comes from overhead cameras and from a pre-defined architecture of the space. We adhere to the tradition of understanding context as a human-centric framework, where spatial and temporal contexts are dynamically mediated by cultural, social, and personal background (Nardi 1996; Dourish 2001). We build on the notion of place and period to operationalize socially defined space and time for our computation of context (Fitzpatrick, Tolone et al. 1995; Koile, Tollmar et al. 2003; Tan, Zhang et al. 2005). Our treatment of space as a structure of relationships is similar to Bill Hillier's Space Syntax Theory (Hillier 1996). Our summarization of events in Tableau Machine is similar to post occupancy evaluation,

which strives to understand the relationship between people and the spaces they occupy (Zimmerman and Martin 2001). The visualizations of the Activity Table (Chapters 3, 4, and 6) mapping motion aggregates to places and periods across rows and columns is similar to (Fleischman, DeCamp et al. 2006).

The Human SpeecHome Project (Roy, Patel et al. 2006) and the Visualization of the History of Living Spaces (Ivanov, Wren et al. 2007) are the two closest projects to Viz-A-Vis in terms of the scope, the goals, the methods, and the resulting system infrastructures. The main difference between Viz-A-Vis and these projects is our fully interactive 3D environment, our localization of pixels and our formal user studies validating our claims about the system's functionality.

Traditionally, activity recognition problems have been grouped into three hierarchies with vague boundaries: low, medium, and high level recognition (Rama Chellappa 2005). Nagel pioneered work to place activity on an operational hierarchy for classification (Nagel 1988). His taxonomy of activity is "change, event, verb, episode, and history." A change is any deviation in a sensory signal, which significantly differs from noise. Classifying changes is a low-level recognition task. Notable systems include (Intille 2004; Philipose 2004). An event is any change pre-defined as a primitive for the construction of more complex descriptions. Classifying events is low to mid-level recognition. A verb describes some activity or the explicit absence of activities (e.g. 'to rest'). Classifying verbs is a mid to high-level recognition. Notable systems include (Munguia, Tapia et al. 2004; Joo Geok Tan 2005). A history is an extended sequence of related activities. Classifying short histories is also a high-level task, for example (Aipperspach, Cohen et al. 2006). Long histories are arguably outside the current activity

recognition hierarchy in that only a handful of systems have tackled it. Recent work exploring this level of recognition is (Huynh, Ulf et al. 2007; Tian, Hampapur et al. 2009).

Bobick defines an alternative taxonomy closer to the standard low, mid, and high recognition (Bobick 1997). The taxonomy defines "motion, activity, and action" as the levels for categorizing particular approaches in terms of their representation and knowledge required to interpret sensor data. Motion recognition emerges directly from sensor data. It is the atomic primitive for vision based activity recognition, requiring no contextual or sequence knowledge. In our general discussion, we include, for example, biometrics, such as galvanic skin response (Sung, Marci et al. 2005), and ambient sensing, such as radio frequency identification (RFID) tags, cabinet switches (Munguia, Tapia et al. 2004), and infrastructure mediated sensing (Patel, Reynolds et al. 2008) as other atomic primitives than motion from image sequences for activity recognition. At the mid-level, low-level features group to form Activities. Activity refers to sequences of movements or states, where the only real knowledge required is the statistics of the sequence. For instance, one instance of the activity "making coffee" may be the sequence of atomic motions "move to coffee machine," "open lid," "place filter," "fill up," "close lid," and "press button." Finally, actions are larger-scale events, which typically include interaction with the environment and causal relationships. For example, the action "get ready for work" may include "make coffee" as part of a longer morning routine and "get ready for work" may also be part of a larger sequence of actions. We will return to this point in chapter 3, when we talk about Tableau Machine and Activity Characterization.

The trend is to classify low-level percepts and compose them into higher-level categories. We argue that the similar to Tableau Machine, the Digital Family Portrait (DFP) (Mynatt, Rowan et al. 2001) is one of the only exceptions to this trend. DFP aggregates motion in the home and maps a single number to the size of icons on a remote display, which serves as a communication trigger between extended families. The aggregates stays close to the raw data and the humans interacting with the system are the units that make sense of the data, not the system itself. The rapid abstraction of DFP consists in the assumption that aggregated motion alone will convey rich and interpretable information about the state of people. Four key aspects of DFP are: (1) it imposes a simple mapping from low-level sensor data to very high abstraction of activity that does not require low-level classification on the data; (2) it is an application-centric approach that puts the problem in front of the solution to determine the level of contextual detail needed; (3) it places the load of interpretation and pro-activeness on the human consumers of the data rather than on the system; and (4) it is one of the few end-to-end applications of context aware computing that has been longitudinally deployed and extensively evaluated.

2.3. Video Visualization and Content Analysis

The motivating principle behind information visualization is that the human vision system is a great parallel image processor. Taken individually, frames from video data sets are immediately processed by a seeing human. The problem with video data sets is the volume that exists today and which currently grows very rapidly. The analysis of video is an extremely time-consuming task. There are a number of video visualization

methodologies proposed (Fels, Lee et al. 2000; Daniel and Chen 2003; Ramos and Balakrishnan 2003; Terry, Brostow et al. 2004; Truong and Venkatesh 2007).

(Ivanov, Wren et al. 2007) present a visualization of the history of living spaces. The authors visualize multimodal (motion detection and video), long-term sensor data that include a number of motion detectors and video cameras. They fuse redundant motion detection data to track walking paths as a low-level perception technique that supports high-level human understanding through visualization. They provide detail through a relatively small set of side and overhead cameras that the user can interactively index through the visualization interface. In relation to our paper, they provide abstract visualization and navigation tools and rapid indexing to original motion sensor and raw video data.

We set similar goals, but present a number of important methodological differences. First, our video data comes from overhead cameras that have a near one-to-one correspondence with architectural space. Second, our goal is to study a broader range of behaviors, more than can be inferred from simply tracking paths. Our main contribution to this discussion line is to explicitly embed the computational perception as part of the information visualization pipeline and discuss the theoretical implications, the challenges, and the opportunities of this methodological shift. Finally, we validate our claims about the system through two summative user studies.

Our general goal is to visualize a multivariate time series in its spatial context. There is a long history of proposed solutions to this task. The most relevant to our work is GeoTime (Kapler and Wright 2004; Kwan and Lee 2004). Kapler and Wright contextualize time series data using the third dimension of a space-time cube that's base

is the relevant 2D map. The main methodological difference in our paper is that we visualize denser data coming from overhead cameras. While GeoTime visualizes one-dimensional paths across 3D space, we visualize two-dimensional surfaces. Kwan and Lee visualize large-scale activity patterns in time-geographies that visualize summarized data for large populations over city-size areas. We visualize spatiotemporally dense data for small populations over building-size areas.

Video visualization is a vibrantly active field of research in recent years. Daniel and Chen present a visualization that holds many similarities to our Activity Cube (Daniel and Chen 2003). They visualize motion in a video space-time cube. They map motion pixels to low translucency in the cube and static pixels to high translucency, thus enabling a human operator to see through inactive sub-volumes of the video cube. Other relevant approaches that model and visualize video as a space-time cube are (Fels, Lee et al. 2000; Klein, Sloan et al. 2002; Bennett and McMillan 2003; Terry, Brostow et al. 2004). Our approach takes these ideas a couple steps further. First, we directly map the video cube to a geographic information system, where the horizontal plane is both image and architectural space and the vertical plane is time. Second, we aggregate motion into regions of interest and linearize the aggregates into the rows of a two-dimensional matrix (the Activity Table) that summarizes the semantics of activity with respect to place and time. Finally, to the best of our knowledge, Viz-A-Vis is the only visualization of video for activity analysis that validates its claims through formal, rigorous, and summative user studies.

The TotalRecall visualization and semi-automatic annotation system shares a number of goals and features with Viz-A-Vis (Roy, Patel et al. 2006; Kubat, DeCamp et

al. 2007). Both TotalRecall and Viz-A-Vis attempt to visualize very long streams of video recorded in real living environments. The two main differences are that TotalRecall visualizes both sound and video and the video visualization is 2D. The video visualization in TotalRecall is not the focus of the interface. In fact, its execution introduces unrecoverable ambiguity between time and space. They slide frames across the screen as cards spread out from a deck across the table. The effect is that each location is a combination of multiple spatial and temporal coordinates in the video cube, thus making it impossible to recover the context of activity directly from the visualization. In fact, their video visualization is not the central focus of their application. It is only a contextualizing tool for the central analysis of speech development in the Human SpeecHome project (Roy, Patel et al. 2006).

Temporal templates present a visual representation of activity as aggregate motion (Bobick and Davis 1996). The nature of the representation of activity in Viz-A-Vis' Activity Cube is similar to temporal templates, with the main difference that motion is spatially contextual. Looking at the layers of aggregate activity across time and space generates three-dimensional maps that have spatial and temporal context. In our approach, we let the human make sense of the sequences.

With MUVIS, (Kiranyaz, Caglar et al. 2003) present a multi-media browser with automatic low-level feature extraction and high-level visual summaries that support navigation, indexing, and querying. The main difference with our work is that they do not contextualize their data in physical space. Their work is primarily concerned with media content and not real-world context.

Our visual aesthetics partially have their origins in the beautiful work of (Larson 1967; Davidhazy 1976; Seale 1995; Sauter and Lüsebrink 1995-2007; Mittelstaedt 2002; Tinapple 2002; Cassinelli 2005; Hilpoltsteiner 2005). Concluding this section, we necessarily mention the main inspiration for many generations of photographers and videographers, the work of Eadweard J. Muybridge and Étienne-Jules Marey (Jaschko 2003). Their work from the 1880s laid the foundation for the photographic capture and visualization of time and the inspiration for the creation of motion pictures.

2.4. Artificial Intelligence and Art

Tableau Machine (TM) is an instance of Expressive AI systems, where Art generates and poses new research questions to the Artificial Intelligence and the AI proposes new artifacts that would have been impossible otherwise. Expressive AI, introduced by (Mateas 2001), proposes that by simultaneously treating design and evaluation issues in Art practices and advances in artificial intelligence as first class research questions, new research agendas are opened in both AI, Art, and Human-Computer Interaction (HCI). Examples of Expressive AI systems are Office Plant #1 (Boehlen and Mateas 1998), Petit Mal (Penny 1997), Giver of Names (Huhtamo 1998), Live Wire (Jeremijenko 1995), and Façade (Mateas and Stern 2003). Examples of generative Art systems are Cohen's Aaron (McCorduck 1991) and Lioret's Being Paintings (Lioret 2005). An approach that mixes perception with creation consists of Art systems that interact with viewers of the work, transforming the viewer into a performer. Examples include Interactive Wallpaper (Huang 2005), Utterback's Untitled 5 (Utterback 2004), and Artifacts of the Presence Era (Viegas, Perry et al. 2004).

In particular, TM is an instance of an interactive Art installation that perceives its environment and reacts in novel ways. The Perceptive Presence work of Bentley et al. is a system built for the workplace that can inform distant parties about the activity of remote collaborators and colleagues (Bentley, Tollmar et al. 2003). The system displays presence and activity via a matched set of glowing ambient lamps. Like TM, Perspective Presence uses computer vision to monitor activities across socio-spatial zones; however, TM uses this information to feed a higher-level interpretative and generative process, rather than directly visualizing this data.

A wider set of concerns, such as enjoyment, aesthetics, wonder, and engagement, are emerging in the HCI community (Blythe, Overbeeke et al. 2003; Gaver, Bowers et al. 2004; Bell, Blythe et al. 2005). Traditional usability design and evaluation methodologies do not directly transfer to these novel approaches. Practitioners have begun to explore new methodologies for design and evaluation. For example, Höök et al present the design of an evaluation methodology for Influencing Machine, a child-like drawing system (Höök, Sengers et al. 2003). One of the authors' main conclusions is that the evaluation of interactive Art systems should help the artists who create such systems understand how users interpret their artifact. Their goal is to give artists a "grounded feeling for how the machine was interpreted and the message was communicated." Note that this is a different evaluation outcome than is produced by traditional user-centered techniques that focus on performance. The usual interpretative model of user interfaces is that it needs to be as explicit as possible about its intended meaning. Metaphors attempt to remain as clear and direct as possible, for example (Blackwell 2006). In this context, interpretation should have only one correct and evident answer. In a domain where interpretation is a

creatively negotiated process between the creator of the artifact and its consumer, the question of interpretation gains unprecedented importance and magnitude. Where there can be possibly limitless interpretations, it is both a potentially enriching and frustrating experience to uncover user understanding of the meaning behind the objet. If anything, it is fascinating (Pousman 2008).

In "Windows and Mirrors," Bolter and Gromala argue for understanding the computer interface both as a transparent window for accessing data and operations and as a reflective mirror for contemplating the medium itself (Bolter and Gromala 2003). They present several examples of explorations of the computer as a medium. The enchantingly beautiful Wooden Mirror (Rozin 1999) and the playfully poetic Text Rain (Utterback 2005) are notable examples of using the computer as a medium for experience, contemplation, discovery, and reflection. We discuss the design and evaluation of Tableau Machine in chapter 3.

2.5. Home Studies

We designed the evaluation of Tableau Machine (TM) as a type of home study where we introduce an external agent into the environment. We investigated how the state of the home before the introduction of TM and we investigated the effects of the introduction. We used qualitative methods to analyze group interviews with elicitation techniques (Mateas, Salvador et al. 1996; Hughes, O'Brien et al. 2000; Crabtree, Hemmings et al. 2002; Bell, Blythe et al. 2005; Gaver, Sengers et al. 2007). A particularly relevant conclusion in (Mateas, Salvador et al. 1996) is that natural activities parse living spaced differently than what the underlying architectural units pre-determine. For example, "kitchen related activities" occur throughout the kitchen, but generally

include the dining room and even the living room. The activity-centric rooms of the house do not uniquely correspond to the architectural rooms of the house.

We also study the home from the Ubiquitous Computing perspective. We install sensors, capture and classify activity, and measure the impact of the technology on the occupants of the home. There is a very large body of work covering this area, for example (Kidd, Orr et al. 1999; Munguia, Tapia et al. 2004; Tapia, Intille et al. 2004; Aipperspach, Cohen et al. 2006; Roy, Patel et al. 2006; Gaver, Sengers et al. 2007).

As we have stated before, our goal was to determine the impact of introducing TM into a real domestic environment. First, we conducted preliminary interviews in order to understand both the culture of the home prior to the intervention and to know how to plan the intervention for its intended effect. For example, we needed to determine where to place the cameras by uncovering the hot spots of the home from the perspective of its occupants. Second, we conducted a number of focus group interviews throughout the deployment to capture data points across time. We were seeking to determine the trajectory of appreciation of the machine (Gaver, Bowers et al. 2004). Finally, we needed to wrap up the study with a holistic image of people's appreciation of the artifact, their games uncovering its workings, their creative interpretations, specially the instances that permitted or enriched reflection, contemplation, and conversation among family members. More than an anecdotal recount, we analyzed the vast data to categorize and synthesize patterns of appropriation (Chapter 3).

2.6. Evaluating Information Visualization Systems

We evaluate two types of analytical metrics regarding visualization of activity. The first metric focuses on measuring the performance and preference of five

predetermined tasks: describing, bounding, searching, counting, and tracking (chapter 5). The second metric focuses on the emergent opportunities to raise valuable discoveries (chapter 6). The performance and preference study measures time to task completion, precision, recall, coverage, and users' task-centric preference. A number of authors have proposed multiple methods for approaching these type of evaluations, for example (Card, Mackinlay et al. 1999; Chen and Yu 2000; Amar and Stasko 2004; Plaisant 2004; Shneiderman and Plaisant 2006). Particularly relevant, (Plaisant 2004) categorizes the types of evaluations based on the tasks, users, and goals. The performance and preference user study in chapter 5 is an instance of a controlled experiment comparing three tools. The study compares a novel technique with the state of the art. Plaisant characterizes the fundamental problem of matching tasks, tools, users, and relevant high-level goals. Furthermore, and relevant to the difference between the user study in chapter 5 and the domain expert study in chapter 6, is the recognition that discovery requires real needs, context, and time. The case study in chapter 6 goes deeper into discovery type questions. While in the user study participants perform predetermined queries, in chapter 6, users raise novel questions and answer them, creating a discovery feedback loop. This evaluation is a qualitative study aiming to determine Viz-A-Vis' capacity to raise discovery of activity patterns relevant to architectural design and evaluation. Notable discovery-focused studies include (Fayyad, Grinstein et al. 2002; Saraiya, North et al. 2004).

In this chapter, we covered diverse work given the multidisciplinary nature of Viz-A-Vis and TM. Next, we describe in detail the design and evaluation of Tableau Machine.

CHAPTER 3

TABLEAU MACHINE: SUPPORTING LONG-TERM CO-INTERPRETATION OF ACTIVITY IN THE HOME

Tableau Machine (TM) is an interactive Art installation that senses, interprets, and generates abstract visual compositions. In this context, we use the Media Theory definition of interpretation. It is a subjective process of negotiated meaning making. TM is a collaborative effort between Dr. Michael Mateas, Adam Smith, Zach Pousman, and the author. Dr. Mateas created the original concept. Adam Smith created the mapping and generating infrastructure. Zach Pousman evaluated TM. The author built the sensing and interpreting infrastructure and deployed the system in real homes. We equally contributed to the design of the artifact and the evaluation. Figure 3.1-a shows two samples of the output of TM and Figures 3.1-b and c show the physicality of TM: a laptop computer, a large screen, a printer, and a number of overhead cameras.

Tableau Machine chronologically precedes Viz-A-Vis (Chapters 4, 5, and 6), the system that is the focus of this thesis. We present Tableau Machine before Viz-A-Vis because the design of the sensing and interpreting infrastructure of TM generated the first version of the visualization techniques that would eventually conglomerate in Viz-A-Vis. In concrete terms, the Activity Table is a byproduct of the design of TM. In this chapter, we will describe the Activity Table in the context of the design of TM and in the next chapter, in the context of the operation of Viz-A-Vis.

The central goal of TM is to engage its cohabitants in a long-term cycle of co-interpretation, where the machine interprets human observable activities, people interpret

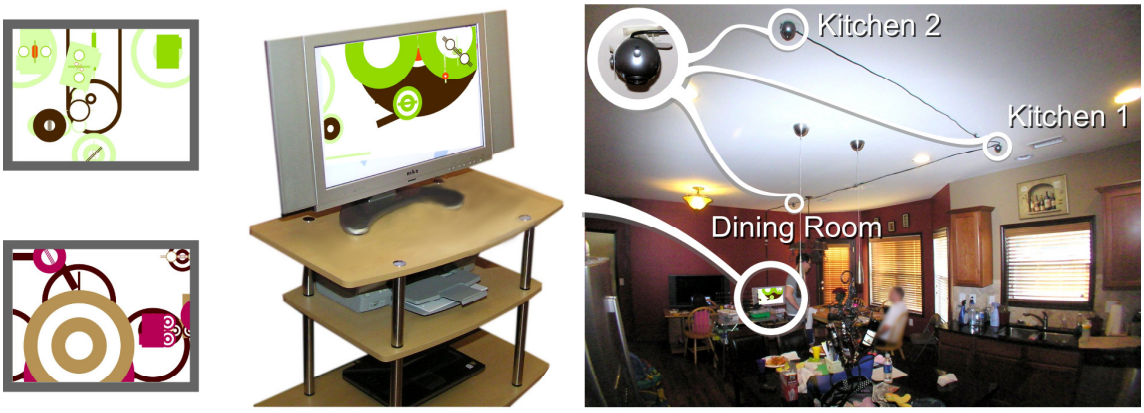


Figure 3.1: Physicality of Tableau Machine (TM): (a) two sample output compositions; (b) large LCD screen, TV stand, printer, and laptop; (c) physical placement of overhead cameras over regions of interest in the public areas of the home, and the location of TM in the center of the image, by the dining room table.

the machine's visual output, and their own activities indirectly. A sub-goal of long-term co-interpretation is to create an aura around the machine that masquerades it as an intentional being, with internal creative states. We aim for the audience to view it as a presence, and not just an appliance. For example, there is a similar contrast between a toaster, an appliance, and a broadcasting radio, a presence. There are two distinctions in this analogy. First, there is nobody actively generating content on the other side of TM. We strive to project this aura of artificial presence through TM. Second, the radio does not react to events in the home. TM does and that puts it more at a level of a pet. As we will cover in the results section, participant treated TM as a pet, giving names and ascribing intentionality and personality to it.

Through co-interpretation, the aim is to open new perspectives into the daily patterns of life that normally remain hidden. Another consequence of co-interpretation is that through highlighting hidden patterns of daily life, TM raises the consciousness of its cohabitants, prompting them to reflect, experiment, and converse on these patterns. In

essence, it is an artifact for meaning making between the authors and the audience, between the artifact and the audience, and between members of the audience. In this sense, it is fundamentally different from other ubiquitous computing or information visualization systems, where the goal is to empower new functionality or to render existing functionality more efficient or effective, as is the case with Viz-A-Vis. Ultimately, the goal of TM is to enrich the perception and experience of daily living.

We made a number of essential design choices in order to achieve TM's goals. First, we embedded *interpretative affordances* in the perception and in the generation modules. An interpretative affordance is a systemic feature that allows the audience of an artifact to negotiate the meaning of the artifact with its creator and, in the case of artificially intelligent artifacts, with the artifact itself (Mateas 2001). Negotiating meaning is a balance between dominant readings and illegibility (Sturken and Cartwright 2001). A dominant reading is one where the author explicitly states the meaning of the message, leaving little room for interpretation. On the other hand, obscuring the meaning of an artifact may render it illegible. Extremely complex or outright random representations (signifiers) will generate illegible messages that receivers will reject (Eco 1979; Shannon 2001).

Second, we embedded *interpretative scaffolds* in TM. While the interpretative affordances increase the possibility of creating free meaning, the interpretative scaffold is a direct-manipulation structure within the system, where the audience can openly observe the effect their actions have on the artifact (Romero, Pousman et al. 2007). The intent of the scaffold is to demonstrate that the machine has a response and to invite the audience to make sense of the more complex interpretative affordances.

With interpretative affordances and scaffolds, we play a design game that has a balance between one-to-one mappings and random chaos. At one end, the reading is *dominant*. The artifact exposes itself completely and does not allow room for creative interpretation. In a dominant encoding, the artifact's meaning is defined by its author with straightforward clarity beyond negotiation. The interactive artifact simply reacts, without intentionality. At the other end, the reading is *rejected*. In a rejected encoding, either the artifact lacks meaning or its meaning is complex beyond understanding. The artifact closes itself to any reading because its behavior is perceived as complete randomness or complex beyond hope (Sturken and Cartwright 2001).

The balance that affords negotiated readings is an artifact that affords creative interpretation. Furthermore, it invites the audience to engage the artifact over a long period. By progressively understanding the interpretative scaffoldings and affordances, the audience learns to make meaning out of the artifact. If we expose the artifact's entire functioning at once, the audience loses interest rapidly, or views the artifact as an appliance and not a presence, with internal states and intentionality.

3.1. Research Question

Consider the overall thesis of this work:

In the process of overhead video interpretation and analysis of activity, combining computer vision abstractions with information visualization techniques provides: (1) improved user task performance measured by time to task completion, precision, recall, coverage, and user assessment; (2) improved user experience measured by user preference; (3) increased user capacity to discover activity patterns; and (4) new

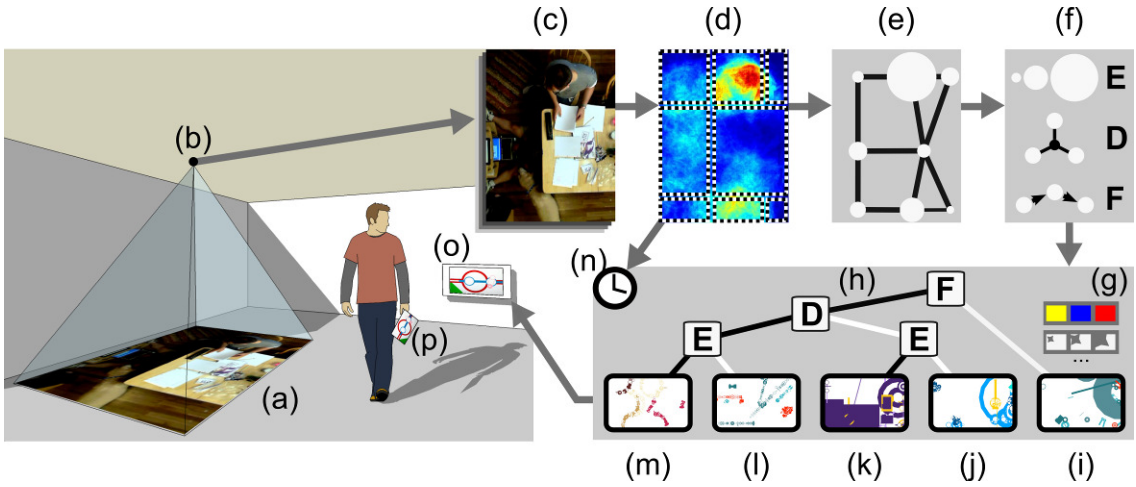


Figure 3.2: Tableau Machine (TM) architecture: (a-f) sensor module; (g-p) generator module; (a) place of interest; (b) overhead camera; (c) image sequence; (d) aggregate motion over place and period; (e) adjacency graph; (f) Energy, Density, and Flow; (g) mapping to color, coverage, balance, and concentration; (h) mapping to a shape grammar tree; (i) NoClust leaf; (j) InnerClust leaf; (k) OuterClust leaf; (l) Kinks leaf; (m) Curves leaf; (n) map of motion to refresh rate; (o) screen display; and (p) print out.

opportunities for creative interpretation, experimentation, conversation, and reflection regarding everyday activities.

In this chapter we address the fourth research question. Can a vision-based visualizing Art installation engage users in a long-term process of creative interpretation, experimentation, conversation, and reflection (Thesis Claim 4)?

3.2. System Architecture

Physically, TM consists of four cameras, a computer, a display, and a printer (see Figure 3.1). Its input comes directly and implicitly from its cohabitants' observable activity captured by the cameras. All the processing occurs in the computer and its output goes to a display screen or a printer. Internally, the processing of TM consists of two modules: the sensing-and-interpreting module and the mapping-and-generating module (see Figure 3.2). TM is a perceptive ambient artistic display. The perception is a function

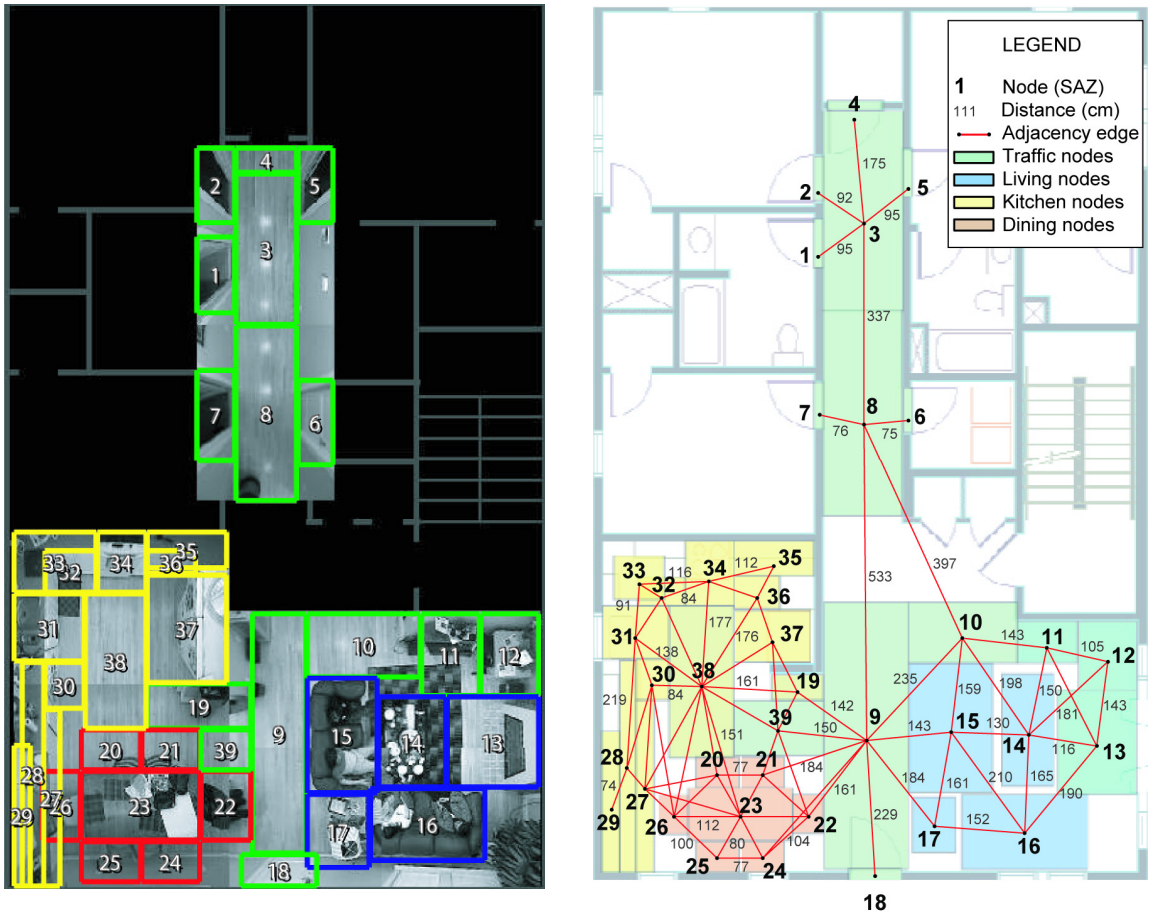


Figure 3.3: Aware Home floor plan, image space, semantic activity zones (SAZs), and adjacency graph. We define the zones manually. SAZs group by room-level regions shown in green, blue, red, and yellow.

of the sensing and interpreting and the artistic display is an emergent property of the mapping and generating module.

3.2.1. *Sensing and Interpreting Activity*

The sensing infrastructure of TM has four increasingly abstract stages: (1) input from cameras; (2) computation of motion; (3) spatiotemporal aggregation of motion; and (4) semantic aggregation of motion into Social Energy, Density, and Flow, three proxies that we define to characterize everyday activity rhythms. We formally define these metrics later in this section and mathematically in Appendix A.

The goal of the sensing and interpreting module is to characterize semantic abstractions of the activity in the home. We want to capture the state of the home; a sense of mood in the atmosphere. We intentionally avoid ascribing concrete meanings to the characterizations for two reasons: (1) concrete and precise high-level activity recognition is an open problem; and (2) not only do we not need concrete and precise labels for this domain, we actually prefer ambiguous and abstract characterizations of activity. This is an instance of an interpretative affordance we embed into the system to increase the possibilities of meaning making. We strive for balance and avoid unrecoverable complexity by still constraining the output with actual activity patterns.

TM senses the environment through a network of overhead cameras. While developing and testing TM, we used the Aware Home infrastructure (Kidd, Orr et al. 1999), with multiple computers receiving the signal from four cameras in the living room, two in the dining room, two in the kitchen, and two in the hallway (see Figure 3.3). In order to create a portable version of TM we used four cameras and a single laptop. We replaced the lens in the cameras to a wide-angle (120°) lens to increase the area of coverage of each camera.

Overhead video readily affords six important technical simplifications to the computer vision problem. First, the camera can be fixed, both in its internal and external parameters, namely focal length, position, and orientation. Second, the frustum is vertical (see Figure 3.2-b). These two simplifications afford a practical one-to-one correspondence between image and architectural space (see Figure 3.2-c). Furthermore, there is a single shallow volume of interest from the ground to people's height. In practice, it is a single plane of interest. Ignoring parallax, the displacement introduced by

perspective projection, mapping pixels to small areas in physical space is a simple, realistic, and robust abstraction.

Third, changes in architectural space (image background) are extremely rare events. We have collected over 13,000 hours of overhead video from 4 cameras in 4 real Atlanta homes, 10 cameras in the Aware Home, 8 cameras in Seagate's Terabyte Home, and 4 cameras in the Bealle Museum in Irvine, California, and the Johnson Museum in Ithaca, New York. In all instances, we observed very little change to the architectural layout, and in the few instances it occurred, it was only minimal and temporal. For instance, we observed small furniture changing location and returning to its original position. The furniture's general position has remained constant in all of our observations. The architectural elements, such as walls, doors, windows, did not change whatsoever.

Fourth, dramatic illumination changes occur very sporadically, typically a handful of times per day. Fifth, the likelihood of people appearing identical to the background is extremely low. At least some part of their body will be of a different color, shade, or texture than the background. Sixth, the likelihood of people holding perfectly still drops to zero very quickly.

Under these real-world conditions, TM computes motion from the original image stream using adjacent frame difference. Adjacent frame difference (AFD) is a robust and well established algorithm in computer vision (Sonka, Hlavec et al. 1999). AFD takes the pixel-wise absolute difference between adjacent frames in the temporal sequence of images. It thresholds the difference and cleans up the resulting binary motion image with the morphological operators open and close. On page 89, Figure 4.2 visualizes AFD.

Next, TM aggregates motion in regions of interest in the image space. We call these regions *Semantic Activity Zones* (SAZ). SAZs roughly correspond to the atomic units of places in an environment, for example, furniture (sofas, tables, and chairs) and architectural elements (doorframes, countertops, and appliances). For TM, we statically defined SAZs by hand using common-sense knowledge of the world. SAZs correspond to places, which provide spatial and social context for activity (Fitzpatrick, Tolone et al. 1995; Nardi 1996; Dourish 2001; Koile, Tollmar et al. 2003).

An interesting future direction for SAZs is automatic and dynamic definition. Automatic and dynamic zone definition grounds the process on the actual behavioral patterns of the occupants and updates it with temporal and cultural contexts affording and constraining behavior. SAZs would enclose semantically different regions much more tightly.

Finally, TM semantically aggregates the accumulated motion in the SAZs. It uses an adjacency graph model to capture simple semantics of space. The nodes of the graph are located at the physical center of the SAZs. The edges encode the physical distance between the nodes. Figure 3.2-e shows a simple version of the adjacency graph. The right side of Figure 3.3 visualizes the graph in detail, including physical distances between SAZs centers in centimeters. From the graph, TM computes three approximate metrics of activity in the home: Social Energy, Density, and Flow(EDF). We defined these metrics. Figure 3.2-f shows a schematic of EDF. Appendix A presents a formal mathematical definition of EDF.

Social Energy is the aggregate of accumulated motion in nodes belonging to sub-graphs of the adjacency graph. The sub-graphs correspond to regions or rooms in the

environment, such as the kitchen and the living room. For instance, during the preparation of dinner, there may be high Energy in the kitchen, low Energy in the living, and medium Energy in the dining room.

Social Density is the ratio between the number of zones active above a threshold of aggregate motion and the distance that separates them, as measured by the length of the minimum weighted spanning tree connecting the active zones. Stated differently, Social Density measures how spatially close activities occur. For example, during dinner, a couple may sit and remain very close to each other, generating little Energy with high Density. In other instances, the couple may be cleaning the entire house, generating high Energy with little Density. Dancers may generate high Energy with high Density and readers in separate locations, low Energy with little Density.

Social Flow is the transfer of motion between adjacent zones in the graph. If the partial derivatives of motion with respect to time in two adjacent zones have different signs, TM computes absolute flow between the two zones with magnitude equal to the positive derivate. Stated differently, Flow tracks the immediate location history of activity. In some instances, physical translation will generate Flow. For example, someone stands up from the living room and goes to the fridge. A trail of Flow will follow that motion across the traversed zones. In other instances, an interchange of motion may generate Flow. For example, if two people next to each other are talking and gesturing, the natural turn taking will generate Flow between the zones they occupy. Again, all combinations of Flow with the other two metrics are possible. There is, however, an almost one-to-one correlation between high Flow and high Energy. Simply stated, high Flow will unavoidably generate high Energy. In that coordinate of this 3D

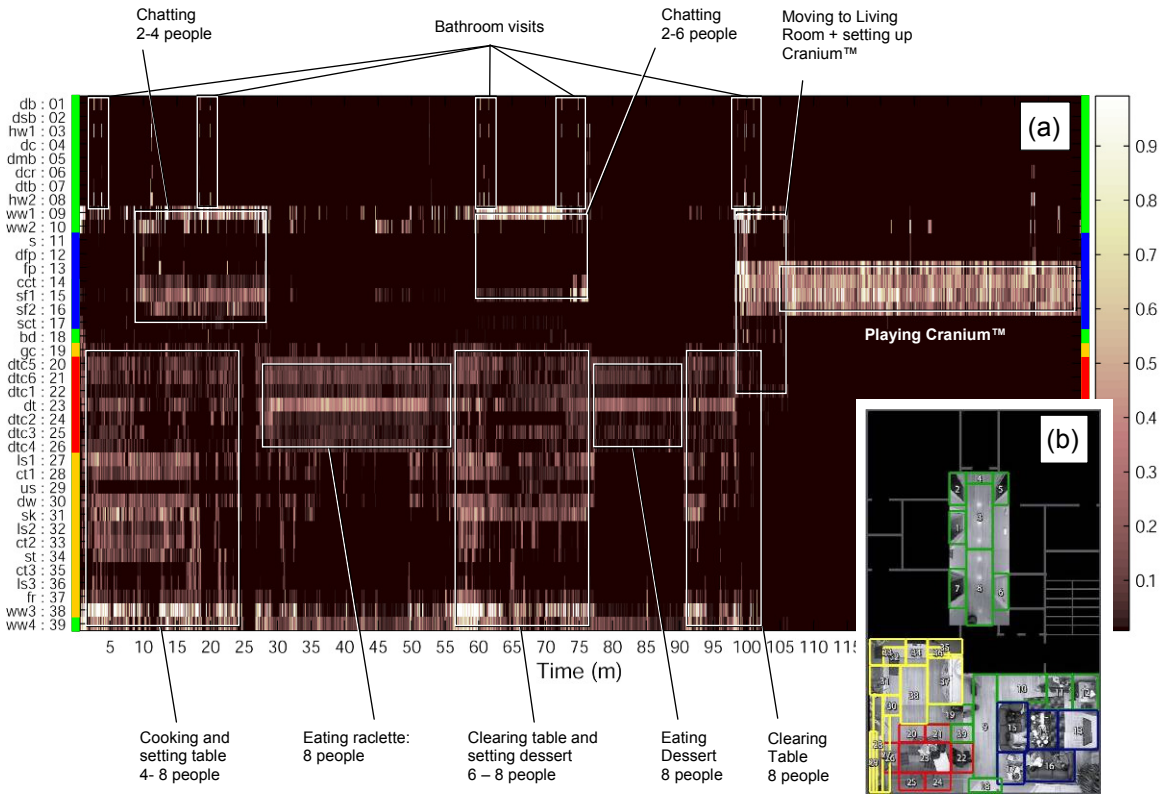


Figure 3.4: The Activity Table (AT) aggregates motion according to floor plan regions of interest called Semantic Activity Zones (SAZs). (a) The Activity Table and (b) the Aware Home floor plan with overhead images and SAZs. AT's rows visualize the level of motion over the places of the home across time, the columns of the table. We map aggregate motion to brightness, scale at right. Because the table encodes spatial semantics, large movements are clearly visible across the table. The data on this table is a dinner party with 8 people. We annotated some episodes during 150 minutes of this dinner party. Figure 3.3 shows a larger image of floor plan.

space, Flow and Energy are redundant. For the algorithmic definition of motion, semantic activity zones, and Social Energy, Density, and Flow, see Appendix A on page 218.

To discover these activity patterns, we conducted studies of the use of space in the Aware Home. We observed groups of participants engaged in natural activity, analyzing several different situations using Grounded Theory methodology (Glasser and Strauss 1967). To facilitate qualitative analysis we built a visualization tool called the Activity Table (see Figures 3.4, 4.4 p.93, 6.16 p.196, and 6.17 p.197). The rows of the table

encode space and the columns, time. The cells are dark when activity is low and bright when it is high. In other words, we map aggregate motion across regions and periods of interest and map the value to the cells of the table. At the right of the table is the scale of level of aggregate motion from 0, black, to 1, white, the least and most observed during this period. We created AT to guide our analysis and programming of models of everyday activity. Our analysis was both qualitative and quantitative resulting in the creation of EDF, our computational proxies of the rhythms of everyday living.

The table includes some detail. The rooms, or regions, are color-coded. Walkways and hallways are green. Living room is blue. Dining room is red. Kitchen is yellow. There are 39 rows for the 39 manual divisions of space, the semantic activity zones. Table 3.1 lists the 39 semantic activity zones. Figures 3.4-b and 3.3 map the SAZ number on the floor plan. Table 3.1 lists the semantic activity zones of AT.

Table 3.1: List of semantic activity zone (SAZ) numbers, abbreviations, and physical places.

SAZ No.	Abbrev.	Place	SAZ No.	Abbrev.	Place
1	db	Door bathroom	21	dtc6	Dining table chair 6
2	dsb	Door second bedroom	22	dtc1	Dining table chair 1
3	hw1	Hallway 1	23	dt	Dining table
4	dc	Door closet	24	dtc2	Dining table chair 2
5	dmb	Door master bedroom	25	dtc3	Dining table chair 3
6	dcr	Door cleaning room	26	dtc4	Dining table chair 4
7	dtb	Door third bedroom	27	ls1	Lower shelf 1
8	hw2	Hallway 2	28	ct1	Counter 1
9	ww1	Walkway 1	29	us	Upper shelf
10	ww2	Walkway 2	30	dw	Dishwasher
11	s	Shelf	31	sk	Sink
12	dfp	Digital Family Portrait	32	ls2	Lower shelf 2
13	fp	Fireplace	33	ct2	Counter 2
14	cct	Center Coffee Table	34	st	Stove
15	sf1	Sofa 1	35	ct3	Counter 3
16	sf2	Sofa 2	36	ls3	Lower shelf 3
17	sct	Side Coffee Table	37	fr	Fridge
18	bd	Balcony door	38	ww3	Walkway 3
19	gc	Garbage Can	39	ww4	Walkway 4
20	dtc5	Dining table chair 5			

We have annotated a few examples of the activity visualized in the table. In overview, it is a five-hour dinner party between eight adults at the Aware Home. They arrive, prepare dinner, eat dinner, clean up, and play a board game, Cranium™. Cranium is a board game with the aim of reaching the end of the board by averting obstacles that require players of a team to perform activities such as singing, acting, sculpting, drawing, computing, verbalizing, recalling facts, and so on (see Figure 5.10 on page 123). Some instances to observe are, for example, that preparing dinner and cleaning up are more spread out than eating and playing. Furthermore, the group splits during the former two. Some people go to the living while others go to the dining room and kitchen. While the first group chatted, the second group prepared or cleaned. Another interesting example is that while people remained generally close both while eating raclette and dessert, in the first instance one or more subjects moved around the space and in the second instance, everyone remained relative put until all were done with ice cream.

Figure 3.4 visualizes these activity patterns. Some activities can be discriminated from others based solely on aggregate motion. For example, “eating raclette” is more active than “eating dessert” even though both are “eating in the dining room.” A Swiss Raclette is an electric grill at the center of the table. On it, people prepare ingredients placed around the table before consuming them. Figure 5.10 on page 123 shows the raclette and the game of Cranium™ that were used that night. Other activities visually different by the paths they leave over the SAZs. For example, people setting up the game board cross multiple SAZ boundaries while people playing the game generally remain in a single SAZ. The same holds for cooking and setting the table versus eating.

To create the EDF dimensions of activity, we first came up with words that could describe what we saw on the table. Summarizing the qualitative analysis, we created the categories *vibrations*, *dispersions*, *conglomerations*, and *translations*. These categories evolved to Energy (vibrations), Density (dispersions and conglomerations), and Flow (translations). Briefly, vibrations are motions that do not incur in a change of architectural location. For example, someone sitting down flipping the pages of a magazine is executing a vibration. Translations are motions that incur in a change of architectural location. For example, someone going from the bedroom to the bathroom, or from one couch to another, will execute a translation. Dispersions are translations that spread people apart and conglomerations are translations that bring people closer. We played around with other terms for the patterns we observed, such as centripetal, centrifugal, periodic, and sporadic. That is the nature of qualitative analysis. There is a process of expansion of concepts followed by a process of grouping and synthesis.

We present a brief discussion of Energy, Density, and Flow. Density is our working approximation of togetherness. We observed that Density is highly correlated with togetherness, but togetherness is not always a consequence of Density. People can be crowded without being together. For example, while clearing the table, people may be crowded between the kitchen and the table, but each person is doing their own task, paying only sporadic attention to others. We can say there is high Density but low togetherness. If we include Social Flow, we can differentiate this activity from, for example, having dinner, which has high Density, low Flow, and high togetherness. Thus, Flow and Density together can characterize the social concept of togetherness better than Density alone.

Similarly, a social descriptor like "busyness" grounds out in our analysis as a combination of Energy and Flow. For example, we observed two instances of high Energy in the living room. In the first instance, eight friends were playing Cranium™. The game has several physical activities, like acting, drawing and sculpting. Furthermore, laughter while playing the game generates a lot of in-zone movement, which accumulates over time and appears as Social Energy. However, in the second instance, when the group put away the game and cleaned up the living room, they also generated in-zone movement and Energy, though the activities are socially very different. Flow helps to segregate the two types of activities. While playing the game people remained within a single SAZ, or had smaller inter-zone flows, occupants clearing the living room moved between SAZs numerous times creating higher Flow values.

However, even with all three measures, many household activities may remain aliased. In other words, they are indistinguishable within EDF coordinates, even if they are very different under common-sense understanding. These activities would produce similar Energy, Density, and Flowmeasures, though the actual activities may be quite different. For example, "dance party" may create similar data to "sibling fight." Rather than attempt to disambiguate these classes, we embrace it as an intrinsic element of the inherent idiosyncrasies of the entity we are creating. In other words, these are the quirks of TM's personality! This mismatch or aliasing contributes to our ultimate goal of providing room for creative co-interpretation of domestic life. For our machine, having a party and having a fight are both instances of highly energetic and dense activity. The goal with our proxy measures is to create intelligible interpretations of activity, while providing room for users to create their own interpretations of system activity. We seek

to give users a novel window into their own activity and a sense of Tableau Machine as a creative presence with its own idiosyncratic interpretation of the household and its internal and intentional states.

Before the final design of TM, we ran a five-hour formative evaluation in the Aware Home with eight participants. The participants found correlations between their activity and the system's compositions, but did not learn exactly how their activities were influencing the machine. This was a positive result; if participants had been able to interpret quickly the underlying mechanics of the system, it would lose its enchanting qualities. Keep in mind TM seeks to engage users enduringly. Enchantment is necessary to sustain long-term engagement. Regarding TM, enchantment is the quality of maintaining the balance between interpretations that are too easy or too hard. It is the zone of proximal development, the enticement to take the next step in understanding, the product of appropriate interpretative scaffoldings. In gaming theory, it is the balance between games that are too hard or too easy, both boring practices.

Here we present two examples from the formative study of how participants A, M, N, and P fabricated simple theories about the system's workings. While participants cooked at the table (a raclette once more), P remarked: "Hey, now it's all red!" pointing to the newest composition. N replied, "It's because I'm burning everything. That explains it!" Later, N said to A: "Ah! Have you seen that?" as she pointed to the composition from the kitchen sink. A replied: "That's pretty! And why did it change like that?" N: "Because we're washing the dishes."

A second spontaneous theory that rose during the experience was a conception of "social balance." After dinner, the women began to clean up while the men retired to the

living room. Female participant A exclaimed “Hey! We had to clear the table.” Male participant M replied, “Well, the system was unstable. If there is too much movement in the kitchen… um [laughter]…I'm concerned [the system] might collapse.” This interaction, while a humorous one, can be seen as a brief moment of self-reflection on a common social trope. It also evidences the type of engagement and stories people may build around the machine.

It was valuable to observe how participants engaged in co-interpretation during the formative evaluation. Participants had socially mediated labels for their activities, such as “washing the dishes,” and had in mind the kinds of representations that might (metaphorically) lead from them. Participants assumed the machine’s compositions had some interpretation of their activity. Though the system output is in fact not based on recognition of specific activities, such as washing the dishes, participants were willing and able to interpret the system output in terms of their concrete activities. This provides evidence that our three proxy measures can support useful and pleasurable readings.

3.2.2. *Clustering, Mapping, and Generating Compositions*

Focusing on Tableau Machine’s interpretation process, we describe how the formal system inputs map to models. TM’s interpretation goal steps beyond simply reducing input bit streams. We designed the system to build representations open to creative interpretation. This goal constrains the space of interpretation processes to those that produce models that are simple enough for the system to demonstrate intention and complex enough to express meaning open to negotiation. Figure 3.5 shows the three-level model of the interpretation.

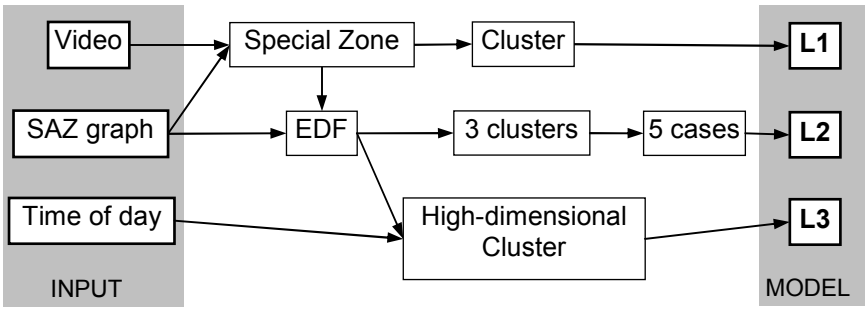


Figure 3.5: Tableau Machine's interpretation module.

We have described the interpretation process to the point of computing EDF. The continuous space of values for EDF over the entire floor plan is still a model too complex to map and generate. Furthermore, individual EDF values cannot represent long-term patterns, which are the system's affordance for long-term interaction with the audience. Thus, we map the continuous EDF volume into a small space of discrete models. An online k-means clustering algorithm maps EDF to states of the home. Because the location of cluster centers continuously updates as the Machine observes more activity, it is a function of the entire history of the system and can begin to address long-term adaptation to old and new patterns.

To support interpretative scaffoldings, a range from simple to complex behaviors, the clustering process operates at varying scales and spaces. This results in different measures, which map to three increasingly complex levels.

Level 1 (L1) bypasses EDF and is solely based on the aggregate motion of the semantic activity zone (SAZ) in front of TM's display. We classify its activity as "high" and "low." Using k-means for two clusters in a one-dimensional space is, in effect, an adaptive threshold. It would be insufficient to fix a threshold at installation time without knowledge of runtime data. This classification affords simple user interpretation of the form "high" and "low." When transitioning to high, people recognize that the Machine

reacts to simple stimuli, for example, standing directly in front of it. The intention behind L1 is to emit the same type of simple information a dog wagging its tail transmits. It transmits acknowledgement and appreciation of presence. L1 creates the first interpretive scaffold of our design. Figure 3.2-n shows the aggregate motion of activity in front of the display mapping to the refresh rate of the generator.

Level 2 (L2) builds explanations of everyday activities using global EDF. Tableau Machine independently clusters house-wide Social Energy, Density, and Flowto produce three "high" or "low" labels. There are $2^3 = 8$ combinations for these labels. We group the eight states into five cases by merging combinations with high global Flow and treating the other combinations as distinct. Figure 3.2-h shows the resulting decision tree of mapping global EDF to five states of the home. The states are: (1) "high Flow" (Figure 3.2-i); (2) "low Flow, high Density, high Energy" (Figure 3.2-j); (3) "low Flow, high Density, low Energy" (Figure 3.2-k); (4) "low Flow, low Density, high Energy" (Figure 3.2-l); and (5) "low Flow, low Density, low Energy" (Figure 3.2-m). The five states afford statements like (high Energy and high Density) "the system thinks the house is active with all the activity together" or just (high Flow) "The system thinks the house is changing states." L2 is the second interpretative scaffolding. It is similar to a dog following direct commands such as "sit" and "fetch." Like the dog, TM needs to learn to interpret the stimuli. Furthermore, once conditioned, TM locks to a state. There is not enough complexity or randomness in this mapping for TM to react surprisingly. Furthermore, it does not convey internal state information or intentionality. It is simply a deterministic mapping input to output. TM invites the user to learn to read it. Once

appropriately learned, there will be no more surprises, just a reassurance of the correctness of the interpretation.

Level 3 (L3) incorporates the regional EDFs and the time of day into a more complex model. Recall that the regions are kitchen, dining room, living room, transit space, and global space. A regional EDF is, for example, Social Energy, Density, and Flowin the living room. Distinct from the other levels, the clustering process in L3 works in a high-dimensional space. Regional EDF contributes fifteen dimensions, three measures times five regions. We map time of day to two dimensions. We view time of day as a continuous angle over a circle. We use the sine and cosine of the time mapped to the 24 period circle. Our intent is to keep midnight close to 1 A.M. Times that are geometrically close together are close together in the clustering space. We randomly initialize and iteratively update 32 clusters in this seventeen-dimensional space.

We ran a data gathering and testing experiment to determine the appropriate number of clusters. We gathered nine days of everyday living at the Aware Home. We will describe this data gathering in more detail later in this chapter and in chapter 6. During our observation of everyday living, we manually segregated over sixty activities, for example, “watching television,” “cleaning,” “eating dinner,” and so on. We grouped similar activities to simplify the clustering space to 32 regions in an attempt to balance irrecoverable complexity with interpretative richness. The active cluster is called the L3 state. While this only affords statements like (cluster-17 active) “the house is in state seventeen,” we designed the space so that these clusters can find their way to common activities. That is, it is possible that the system could behave in a manner consistent with statements like “The system can tell we are sitting down to watch our favorite television

show," but only because it has a model of what regional EDF looks like during the show's running time. However, if no interesting patterns are discovered, it is easy for the audience to dismiss the resulting behavior of the system as "just more randomness."

This reception would be an instance of the rejected reading we discussed earlier. It would signify users are incapable of reading the complex encodings and are dismissing them as random eccentricities. Nevertheless, we argue that is the natural interaction between complex entities. Prior to deeper cultural and personal understanding, numerous behaviors seem arbitrary or contrary to one's beliefs and customs. People require long exposure to develop this level of understanding. Returning to the analogy of Tableau Machine as a pet, L3 communicates personality-based actions and reactions. It is the set of traits that make each dog unique to its owner. For example, L3 expresses liveliness, patience, irritability, and empathy, especially when interpreted by human observers, who routinely anthropomorphize objects and animals that exhibit behavior. It is a goal achieved not by TM alone, but as a product of co-interpretation, the machine interpreting the human and the human interpreting the machine and self-interpreting through the machine.

L1, L2, and L3 support behavior interpretation with different levels of complexity and ease-of-explanation from outside of the system. L1 updates quickly in response to audience provocation in the special region in front of the display. L2 responds to global activity more slowly. Finally, L3 responds to recognized patterns only over very long periods (as cluster centers adapt). Each level corresponds to beliefs the system has about its environment arising from its idiosyncrasies. The cumulative effect of these idiosyncrasies form the sense of personality and intentional presence we strive for.

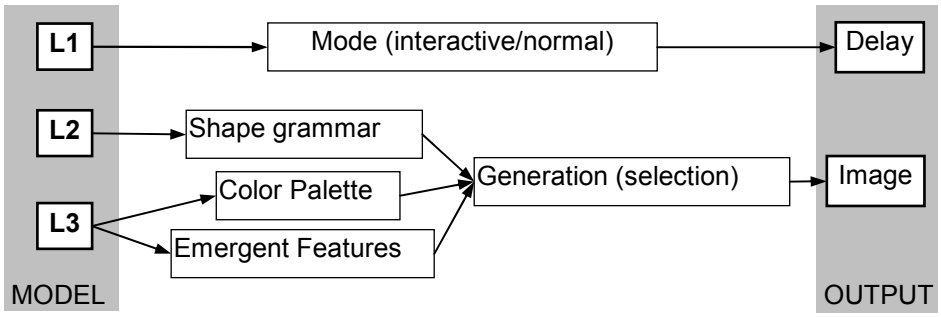


Figure 3.6: Tableau Machine's expression module expressing the state of TM's beliefs.

Now we move to the generator module of Tableau Machine. The generator module of TM is less complex than the interpretation module. Here we will describe how the space of L1, L2, and L3 states map to particular images. Figure 3.6 shows the TM's expression module.

L1 is the simplest model produced by the interpretation process. We map it directly to a simple, visually prominent output. When the L1 state reads high, TM goes into "interactive mode" where new images are selected very quickly, causing the display to fade from one composition to the next after roughly one second. When the system is in "normal mode," it selects a new composition about once every two minutes. We designed this behavior so it has a way to say, in its own language, "I acknowledge your presence and you are engaging me." Here we have mapped a simple belief to a simple output, avoiding leaning on connotations that the artifact does not understand. We adopt this "impedance matching" heuristic to avoid losing the audience because the viewers may read too much meaning into the observed outputs of the system. By "impedance match" we mean maximizing the input-output encoding of meaning. Furthermore, we avoid hiding too much of the system's potentially interesting interpretation from the audience.

TM's L2 puts a basic requirement on the images selected for display. For each of the five L2 states, our mapping dictates that a specific shape grammar be used to generate

the output image, prescribing a distinct visual style. Figure 3.7 on page 54 displays images from the five shape grammars. A shape grammar is a production system consisting of compositional rules that generate valid sentences in its visual domain (Stiny 1972; Stiny 2008). Similar, for instance, to the natural language grammar which generates valid sentences in English, with syntactic and semantic structures, shape grammars generate valid sentences in the compositional domain defined by the author of the grammar. With shape grammars, forms can be instantiated that did not exist previously. Formally, a shape grammar consists of a vocabulary and a set of production rules. The vocabulary includes a start and an end state, a set of non-terminals and a set of terminals. The non-terminals determine the structure of the production, while the terminals determine the detailed look. In the five productions in Figure 3.7, the non-terminals determine, for example, the straight compositional lines of "kinks" versus the curved lines of "curves." The terminals, on the other hand, determine the actual elements that are visible, for example, circles and squares.

Though there are aesthetic reasons for mapping certain L2 states to certain grammars, the system is only aware that distinct L2 states are represented with distinct grammars. The connotations encoded in the mappings are simple. For example, when Flow is high, there is no structure, no clusters. The five mappings will receive treatment soon. The point here is that to attempt to embed any more meaning than this weak connotation in the mapping of grammars would push on the impedance matching heuristic because the system would "use words it does not understand." Stated differently, we restrain the conceptual level of the abstractions to avoid producing gibberish. TM produces simple statements that, in fact, have at least one interpretation

from which real events can be spotted. TM does not produce complex statements randomly without regard to real input and production rules. Again, our goal is to maintain the balance that has the potential to elicit creative interpretation.

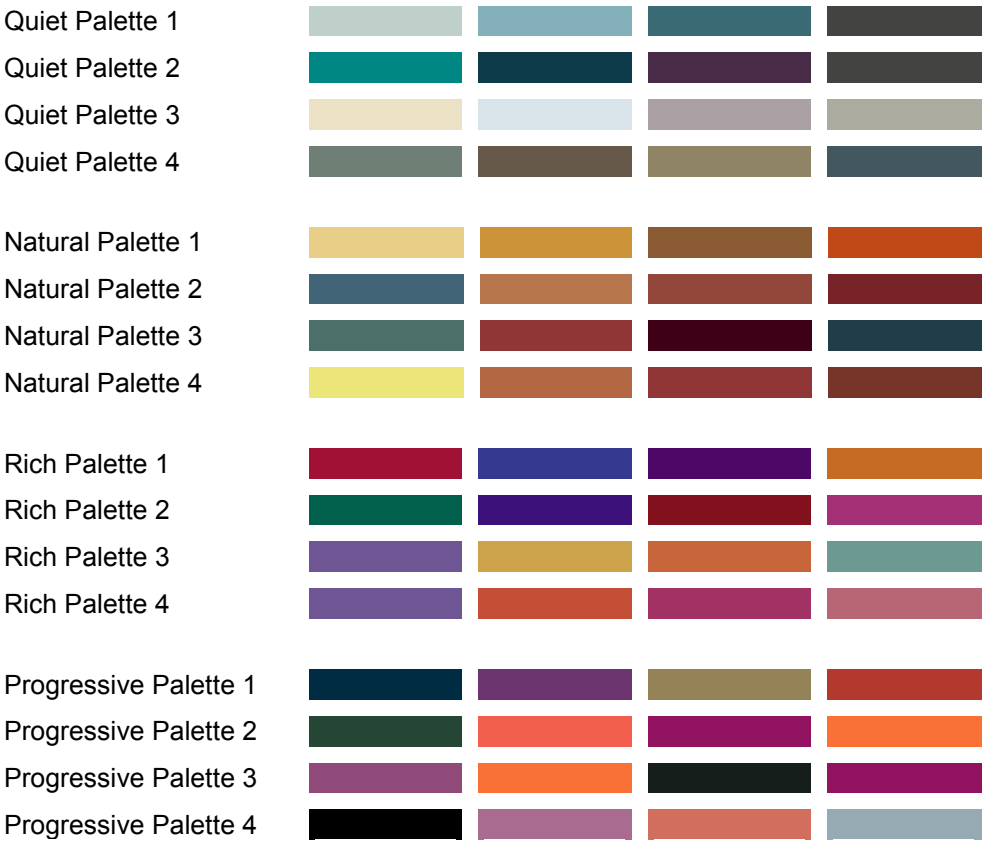
The shape grammar does not dictate the entire appearance of the final images displayed. We have included a level of constrained randomness for providing the details that differentiate images we sample from each grammar. The L3 state of the system directly controls the distinct emergent visual properties of an image as well as which color palette embellishes its design.

For aesthetics, we used a graphics design color index to choose the color palettes and families. Table 3.2 shows four *families* of color with four *palettes* each. The families are *Quiet*, *Natural*, *Rich*, and *Progressive*, categories described in (Krause 2002). Combinations within the same family should be perceptually close to each other, while as distant as possible from combinations from the other three classes. At the same time, combinations within a class should vary enough to give a sense of variety and novelty, while the evocativeness remains roughly the same, similar to synonyms in English.

The Quiet palette consists of pale, dark, and pale and dark color combinations, which can be calming, serene, and relaxing. Hues in the blue, blue-green and blue-violet spectrums convey a visual quietness to many people. People do not usually see these colors as signals for alarm as may be the case with red, yellow, or orange tones. A muted palette can add a sense of calmness or nostalgia to a set of hues. A group of dark hues with minimal difference in value can impart a quiescent feel, perhaps edged with mystery and an underlying tension. A combination of extremely pale hues used throughout a piece can also be useful for establishing a low-key emotion. The Natural family of palettes

consists of colors borrowed from the earth, sea, and sky: brown, tan, and red tones of soil; greens, yellows and oranges of foliage; blues and reds of sky and water. Designers trying to address viewers looking for a straightforward presentation, often use these colors. The Rich family consists of combinations that contain hues used historically to convey affluence, honor, royalty, tradition, and wealth. Violets and deep blues are often combined with full shades of green, gold, red, and maroon. The Progressive family combines hues and combinations that are among those that seem to continually resurface in contemporary media that proclaims a trend-setting message.

Table 3.2: Color families Quiet, Natural, Rich, and Progressive and four four-color palettes per family.



In addition to the visual style imposed by the grammar, the L3 state prescribes the three emergent compositional properties of "coverage," "balance," and "concentration" of images. *Compositional Coverage* determines how much the foreground covers the background. *Compositional Balance* describes whether the foreground detail is left-heavy, right-heavy, or balanced. TM only considers horizontal balance. *Compositional Concentration*, similarly, describes whether the foreground detail is center-column-heavy, side-columns-heavy, or not distinctly one way or the other.

The point is that each of the 32 distinct possibilities for the L3 state maps to the presence or absence of three different compositional features and to one of four color palette families.

The result of the expression process so far is a grammar name, a set of visual features, a color palette, and the update rate from L1. A single step remains before the output is ready. We turn these requirements into a concrete image using the generation component of TM. The output image appears on the display or the printer, where the audience is free to "decode" it. In the next section, we will look at the generation component in more detail.

The expression component, as a whole, translates a simple set of beliefs into a personality with its own language. Where possible, we designed the expression component void of a clear association between activity and system behavior with enough complexity to wash away possible mistakes and misinterpretations. Recall that we aim to support and engage the audience in creative interpretation, not to have the system declare its own interpretation of the environment to be the general truth.

For the system to support creative interpretation, TM must display images that are engaging enough to spark investigation and the properties controlled by the L2 and L3 states must be distinct enough to be noticeable. These constraints, intersected with the authors' aesthetic motivation to maintain a single style with a single voice, set high expectations for the generation component.

Ideally, we would have a simple method to synthesize an image each time the system expressed its state. After exploring other venues, we chose a generate-and-test method for achieving our combined goals. The "generate" part uses shape grammars to over-produce a space of images that includes those we are looking for as well as many others. The "test" part uses basic image processing to assign labels to images that we can use to filter out only appropriate images. Taken together, these independently controlled parts of TM's generative module comprise a high-performance and parameterized image synthesis process.

We generate images using context-free (Chomsky Type-2) shape grammars (Stiny 1972; Stiny 2008). Specifically, we use Coyne's open-source CFDG (Context Free Design Grammar) package (Coyne 2005). Informally, shape grammars in CFDG are sets of simple rules describing how to draw shapes in terms of other shapes. Rules build up in terms of primitive (terminal) shapes such as circles, squares, and triangles, in terms of other rules, or in terms of themselves as is the case with recursive rules. Furthermore, several rules may share the same name, indicating that there are several ways to draw the named shape. This practice yields a non-deterministic grammar. This non-determinism, coupled with exploring a large space of random seed values, is what gives rise to the immense space of images that we select from in the generation component.

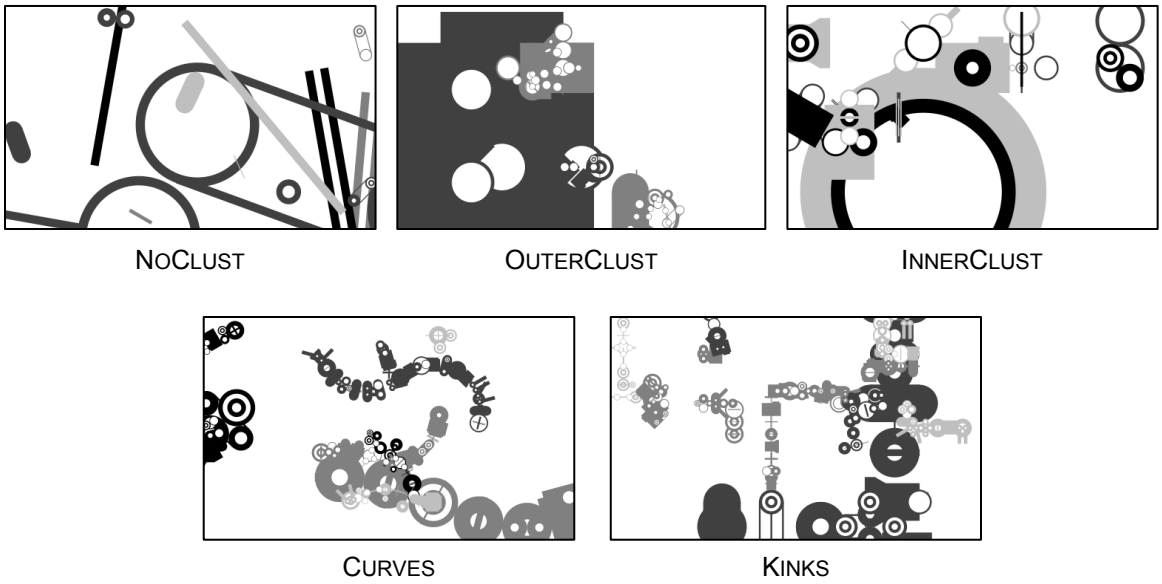


Figure 3.7: Example images from the five shape grammars. Color assignment is a separate process.

The main shape grammars consist of a stack of grammars and a shared library of TM-specific shapes to reduce complexity and enforce a common visual motif. The opposing goals are to maintain uniformity on the one hand, and uniqueness across productions on the other. Figure 3.7 shows a sample from each of the five main grammars. Each grammar contains rules that describe the overall placement of grammar-specific non-terminals, for example, straight lines for Kinks and curves for Curves. The grammars also contain the definition of the non-terminals in terms of the rules from a grammar called "ANY" that describes the terminals common to all of our main grammars. We present a brief description of each main grammar.

The simplest grammar, NoClust used for high-Flow L2, haphazardly scatters ANY shapes, making use of minute angular offsets and gross scaling to give a disheveled look. The goal of NoClust is to connote lively disorder, similar to a household late for school. Again, we avoid any further meaning ascription.

The condensed-looking grammars used for high-Density L2, Curves and Kinks, produce worm-like shapes composed of long chains of ANY shapes that flip orientation and heavy-versus-light along their length according to an improvised, first-order Markov model. Kinks and Curves activate when there is low global Density. Kinks connote directed busyness and curves, wandering strolls.

The gaseous-looking grammars used for low-Density L2, OuterClust and InnerClust, are populated by composite clusters. In OuterClust, clusters flourish by growing a seed shape and surrounding it with smaller, similar shapes of the same color, creating a bubbling silhouette of terminals. In InnerClust, clusters thrive by nesting shapes of differing color at increasingly smaller scales inside of an outer shape. Both InnerClust and OuterClust denote a densely packed home with and low and high activity, respectively. They connote the idea of togetherness. InnerClust connotes togetherness calm unity. OuterClust connotes lively unity.

Each grammar describes images that are distinct from other grammars; however, within each grammar there is still an effectively infinite space of variation. Furthermore, between grammars there is still a sense of unity. The arrangement of shapes in a final image is the result of sampling from the generative space of a shape grammar using a specific seed value for the internal, randomized rule selection processes.

The rendering system for our shape grammars is capable of producing high quality and full screen images in a vector (shape-based) image format. While we use this format for display, we will see that we will have to generate raster (pixel-based) versions of the compositions to support automated analysis of their content in the “test” process.

Recall that the expression component mapped the L3 state to distinct visual properties of images, not just to the name of the grammar that generated it. In order for the system to have a better idea of what the compositions it generates "look like," we pass low-resolution, raster images to an image analysis program. This program looks only at very basic properties of a foreground-background map (independent of intensity). This process results in numerical assignments for the three visual features mentioned in the expression component (coverage, balance, and concentration). These values are in a continuous space, so we manually chose a working threshold and used it to assign categorical labels for each feature. In other words, we manually chose what is good enough from a sequence of samples and from there let the system run independently.

A pixel-level analysis of the images is important because many visual properties are not obvious from a shape-level description of an image. For instance, an image with a single large shape obscuring many small ones appears to be a very simple composition at the pixel level, however the shape-level description would suggest a complex result. Alternatively, if all of the shapes in an image happen to cluster together on the left half of the image, the viewer may perceive a distinct imbalance that is, again, not obvious at the shape-level.

When given enough time, our generate-and-test process can produce an image suitable for expressing any state of the system. However, by design, TM is a soft real-time system that depends on meeting deadlines to support live human interaction. Because of this constraint, we run the generate-and-test process off-line, save an overproduction of the results, and let TM pick unseen instances at runtime.

Because the soft real-time parts of the system only require that an image be produced with a set of requirements coming from a finite space, we can pre-compute a large set of images for use in each state of the system. In practice, this meant sampling about 26,000 images from each of the five main grammars. After analysis, we discard the bulky raster version of each image. We saved a compressed version of the vector source for images along with the result of analysis in a database and used the database in read-only mode for live installations. Thus, the deployed system need not include the ability to sample from a grammar or analyze images, greatly simplifying the software aspect of our live installation.

Clearly, in terms of the visual output of the system, whether the generation process occurred in the home or in the studio before installation is not important. This part of the process does not learn from experience. In either case, live TM produces suitable images without repetition during the three eight-week installations at participant home. We will discuss these evaluations in the following section. In terms of the interactive nature of the output, our choice was critical, as, for a given set of requirements on an image, it may have taken hours of search to find a suitable image in the generative space of a shape grammar.

For Ubiquitous Computing systems that have interaction patterns that are not known in advance, a longitudinal evaluation is required; we cannot know in what ways (or even if) households will appropriate technology until they have it long enough for appropriation to take hold at the day-to-day level. Gaver et al. call this a “trajectory of appreciation” (Gaver, Bowers et al. 2004), a timeline of the degree to which householders engage and use a technology device over a long period.

Finally, to make our lookup table more effective, we restricted the generate-and-test process to grayscale images and left the computationally simple final application of color palettes to the on-line system. Thus, our database-driven image synthesis process can use any of over 2,000,000 unique and high-detailed compositions. Note that only about 50,000 images appear over the lifetime of a TM installation, most going unobserved. The practical result of this process is that participants do not see the same image twice. This aspect of TM had important implications for the printing practices at each household, as we will see.

The final step is to render continuously new compositions on the large display and to print the images on demand. We programmed a simple interface for one button printing at any time during the deployment. We will revisit the printing practices in the evaluation sections.

3.2.3. *Testing and Calibrating*

Before we describe the evaluations, we present a brief recount of the testing and calibrating of Tableau Machine (TM). Testing and calibrating TM was an extremely complex task. To test TM, we deployed it in the author's house for six months (over 4000 hours). During this time, the author and his wife had a daily exposure of at least 90 minutes to the renderings. The display was located in the dining room, the public area of the house with the highest daytime occupancy. With an average generation and refresh rate of 2 minutes, we consumed and interpreted over 8,000 unique compositions. In some instances, we controlled our behavior to influence the machine and predict its outcome for over 600 productions. TM has a number of sensing, learning, and rendering parameters that took non-trivial time to test and tune. While the typical turn-around time

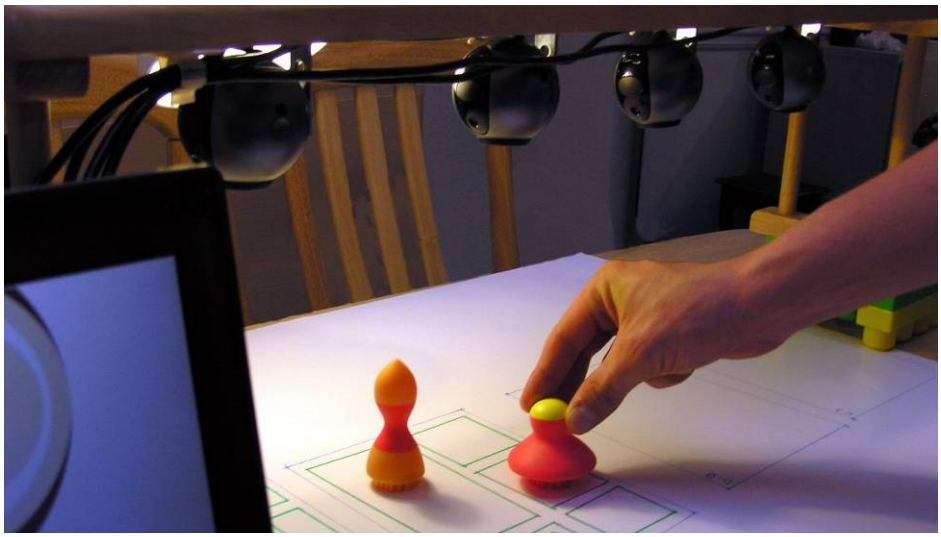


Figure 3.8: Testing Tableau Machine with a scale model and cameras.

of a test-and-tune cycle is a few minutes or, at most, a few hours, TM's test-and-tune cycle ran for a few days. In other words, we would tweak a parameter and wait for two to three days to appreciate the consequences. The myriad of parameters include motion thresholds, activation thresholds, learning speed, weights in the mapping attributes, level of randomization of rendering structures, color palettes, and coverage. During the six-month period, we had guests 14 times. Every time a guest arrived, we introduced them to Tableau Machine and invited them to play with it. To experiment with the system's response to their behaviors. We gradually explained how the sensing worked and let them continue their experiments. Eventually, for recurrent guests, we explained the entire workings of the piece. It was an interesting game for them to see the evolution of TM across time. Their comments and suggestions ranged widely. The most common suggestion was to include animation. Early in the design cycle, we decided to exclude animation because we did not want TM to feel like a screen saver. The second most common comment was the actual recognition that it was tracking motion and mapping it to features of the composition. People read the renderings as intensity of motion under

the cameras. The readings never got to the point of Density or Flow, or togetherness or busyness. They were basically, "now it's changing fast because there is a lot of activity here [in the dining room]." This is the first level of interpretative scaffolding we included in the machine through the L1 mapping between activity in front of the display and refresh rate.

In order to reduce the test-and-tune cycle we devised a scaled model of the home. We reduced the size of the home and accelerated the passage of time. This scale model allowed us to test-and-tune 20 times faster than with real data. Figure 3.8 shows the scale model. Although this model helped in some instances, most of the tuning occurred in the real setting. When we finished tuning and testing the robustness and appropriateness of TM, we moved on to the user studies we describe next.

3.3. Deployment and Evaluation

In order to explore the degree to which our design choices matched our goals of supporting and encouraging co-interpretation, we deployed TM into three homes in the Atlanta area (Figure 3.9).

3.3.1. *Procedures and Methods*

We recruited three homes from posts to a local online message board (http://www.craigslist.org). TM "lived" for six to eight weeks in each home. We compensated $500 the families for their participation, their energy consumption, the use of the space by TM, and the participants' time, especially during interviews.

We selected the households based on a variety of factors. We intentionally chose households that contained children living at home. We were not making normative claims about households. Rather, we chose families with children because they have two

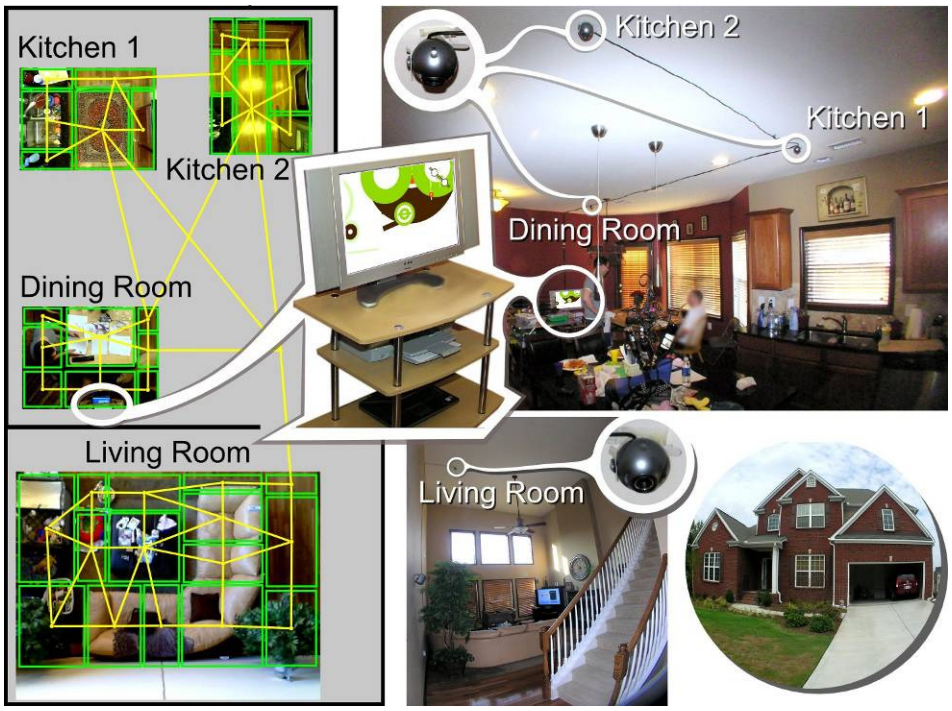


Figure 3.9: Typical TM installation (House A). Floor plan, overhead fields of view, semantic activity zones, and adjacency graph on the left.

attributes of interest. First, we speculated that households with children would generate a greater dynamic range of activities. By dynamic range of activity, we mean the variation from quiet to loud and from calm to frenetic. We were interested in finding homes with numerous social states and configurations from calm to frenetic. Second, we were interested in feedback from a wide population of users and we felt that children might engage with TM differently. We describe households A, B, and C in Table 3.3 below.

3.3.1.1. An Unexpected Pseudo-Control Condition

Household A received a TM that was different from the others. TM appeared to work even though we miss-configured the system in household A. The camera data was captured and images were created. However, the only data from the cameras read by the generator module was the activity coming from the special zone. The fifteen Energy,

Density, and Flowpoints where disregarded because of a communication failure between the interpreter and the generator. The generation system read from an empty file, and produced images influenced only by the computer's clock and the motion in the especial zone. The images were severly underconstrained, leaving too many parameters to chance.

Table 3.3: Describing home demographics and deployments. Names have been changed to protect identity. Householder's first initials match household code (e.g. Byron lives in household B).

	Household A	**Household B**	**Household C**
Parents	Andrew 40s (technologist) Amanda, 40s (stay-at-home mom)	Byron, 40s (co-owner of daycare with spouse) Betty, 40s (co-owner of daycare with spouse)	Carla, 30s (translator)
Children (full time)	Alice, 10 Andy Jr., 5 Amy, 7 mos.	Brian, 14 Brianna, 10	Charlie, 10
Pets	Large dog	None	Two cats
TM	Partially functional (see below)	Fully functional	Fully functional

We did not discover this technical problem through the evaluations because TM generated images that superficially made sense to us. When we visited the home as evaluators, the images seemed to be making sense, though it was hard to determine due to the unavoidable short exposure we could have with TM in housholds. Part of the problem was that the clustering seeds moved into locations spurred by the clock time, giving the impression of learning and adaptation. We found the communication breakdown on the final week of the deployment at Household A, just as we were beginning to dismantle the system. Our interviewer did not have any inkling that something was technically wrong, though he did find that the family was less engaged with TM. We discuss this in the

following section. Household A deployment serves an important check against one potential problem with studying systems designed for open interpretation. What happens when you present random ambiguous productions to people versus real ambiguous productions? Will people equally make sense of the two given their natural propensity to find patterns? Will people, when shown random or pseudo-random outputs, proceed to deeply interpret and engage it?

If the human participants, without active interpretive support from the system, are doing all of the interpretative work, the system is in some sense a failure. This result would indicate that a simple random number generator could replace the sensing module and the system would still have similar impact on people's appreciation of it. In the case of Household A, participants did not do so. Though this is not an experimental control by any means, the differences amongst these two conditions (broken TM versus a correctly functioning one) were real.

3.3.2. *Interviews and Elicitation Techniques*

Our investigation was qualitative. We sought rich accounts from family members about the rhythms and activities of the home, both in pre-installation interviews and in weekly interviews at the home. We recorded the interviews with a tripod-mounted camcorder. We transcribed and analyzed this data. In addition to qualitative interviews, we used a small set of elicitation tools during the interviews to support both retrospective stories of home life and reflection and conversation around TM.

- A feltboard is a tangible representation of a set of objects or interface elements. The feltboard consists of a schematic floor plan of the environment. The pieces on the board consist of people and objects in the home. Participants can move around

Figure 3.10: Interviewing a household and discussing TM printouts (left). Householder selecting words in the word game (right).

pieces on a board, either individually or as a group, to tell stories or to design a configuration (Rode, Toye et al. 2004). We built a feltboard representation of each home. We created tokens for furniture, including a token for TM, and little figures representing the householders. The feltboard improved recollections of activities and recounting of anecdotes. It materialized, contextualized, and provided visual focus to the discussion, giving participants a wider range of vocabulary and gestures to communicate their daily living experience.

- We deployed a word game with householders near the end of the evaluation. We presented a large set of words; each printed on a strip of paper, and spread them out before the family. Each individual selected a small set of words that best described salient aspects of TM. It included words about the physicality of TM ("screen," "camera"), the productions ("circles," "lines"), metaphorical ascriptions ("blender," "thermometer," "mirror"), as well as judgments ("boring," "engaging," "befuddling"). The word game also included freeform blanks for householders to fill in (Figure 3.10 – right).
- TM Printouts became one of the most salient mechanisms for getting feedback from householders. Householders could easily print images they liked (or wanted

to write on, or for any other reason) via a small keypad. Often, in a week, families printed a stack of images and would spontaneously suggest that we go through all of the prints they made (Figure 3.10 – left). We will continue with the analysis of printing below.

3.3.3. *Analysis*

From the interviews, we gathered voluminous qualitative data. We created large categories of themes that emerged from the transcription and analysis of the video recordings of the interviews. The categories are: trajectory of appreciation, experimentation with Tableau Machine's inputs and outputs, the endpoints of the trajectory of appreciation, hints of TM's personality, printing practices, deepening of reflection, and feelings of being watched. We discuss these categories and their relationship to co-interpretation.

3.3.3.1. Trajectory of Appreciation

Gaver et al. describe the process of adoption of a new technology as a trajectory of appreciation (Gaver, Bowers et al. 2004). At the outset of the trajectory of appreciation, users embrace a new technology merely because it is novel. Unavoidably, these novelty effects wear off, as the realities of use (functional limitations, fragility, and problems) become apparent. These obstacles typically bring appreciation below the level prior to the introduction of the technology, a state of anticipation. This happens when a task or activity becomes more difficult and less satisfying by the introduction of the device or technology. After weeks or months, the technology settles into a steady state of use. Either the technology is rejected, or it is adopted, and usage and satisfaction rise. Users find ways to route around the difficulties they might face and to use the technology

in spite of its limitations. A longitudinal analysis of appreciation is particularly applicable to TM, since TM cannot be fully understood, much less appreciated, in a few minutes of observation. As such, watching the trajectory rise and fall is instructive as to how well or how poorly the system becomes integrated into the lives of householders.

3.3.3.2. Experimentation with Tableau Machine Inputs / Outputs

Householders universally began with a sense of puzzlement. Brian stated "How can you tell what's what?" (Note: Names have been changed, but each name's first initial matches the household code, so Brian, a 14-year-old young man, lives in household B). Andrew, a professional engineer claimed at an early interview that he had performed an "experiment" with TM.

Andrew: "No pictures are ever the same. I've already tried to do that." Interviewer: "Is that right? You tried to do that? How did that work?" Andrew: "... So yeah, no, there was no one else in the house. So I just sat right here and I was just [he mimes holding still], and I was looking at it and looking at it and... But I could never get it, the same exact image. ... Well that's what I'm trying to figure out, If it's taking numbers, images, from [the cameras], then why is it not the same picture every time [while nothing changes]? You know?"

With TM, even if the system inputs are the same, different images emerge, though they will share the same style. This would not be apparent after just a few days. While none of the B or C householders were scientists or engineers, they also reported performing experiments to figure out how TM works. Some of these experiments were individual activities while others took more than one person to perform. Early on, Carla wrote on a particular image, "Question: I wonder if different colors in the rooms -- in

front of the cameras will produce different colored art work? Something with different motions -- harder to quantify, I think."

3.3.3.3. The Endpoints of the Trajectory of Appreciation

In household A, the logic by which TM mapped home activity to images was impenetrable, since there was no such logic. Householders A were proud of their discovery that TM had patterns of colors and compositions that differed between mornings and evenings. They claimed that certain compositions were "morning ones" and others "evening ones" based on colors and composition attributes that were influenced by the computer's clock. However, family A did not draw deep meaning from the relationship between morning and evening images and the differences in the home at these times. In contrast, households B and C, with correctly functioning TMs, stayed quite engaged with TM throughout the full six-week deployment.

Household A, in the word game, selected simplistic words to describe TM and were neutral as to their experience. Alice even selected the word "stupid" to describe TM, while other family members were more charitable. Amanda selected "confusing" amongst other words.

Household B stayed engaged with TM. They found that the productions were very much about the family. Near the beginning of the deployment, Betty (the mother) began to describe images as being views of the house, either from above or from other perspectives. Other householders followed along in this reasoning, and pointed out clusters that were "the kitchen table" or "the hallway." As the deployment progressed, B householders began also to see individuals in the images, and to draw parallels between activities (such as a boisterous dinner) and the images (a large round shape full of messy

Figure 3.11: Tableau Machine printout on Household B's fridge. Participants interpreted the image to mean "the smiling face [in profile] of the father while cooking."

shapes on top, including a set of lines that formed something resembling a fork). The family was quite enamored with this image, and others that represented moments around the house. In the last week, Betty found an image that looked like a smiling face, which she took (or pretended to take) as an image representing her husband cooking at the stove. At the interview, she was very proud of the printout and asked if she could keep it. She hung this picture on the refrigerator. Figure 3.11 shows the printout displayed on the fridge.

Household C ended with feelings of intimacy toward TM. Our impressions come from two occasions where householders did something special with TM. As we uninstalled TM, we noticed a particular image that had been written on by Carla. She smiled, was very excited about the printout, and asked if she could keep it. Carla wanted to keep all of TM images. We asked what she might do with them. Carla replied that she wanted to make a photo album of TM images, alongside images of the householders. She

said, “You know how you put in pictures of a vacation? These will be pictures of when Niko [their name for TM based on the brand of the LCD screen] was living with us.”

There are two interesting parts of her answer that give us confidence that she was creating more than a casual connection to TM. First, the household’s experience with TM is memorable and positive enough to warrant the investment of time and money to make the photo album. Second, she described TM system as having a social presence in her home. The family felt differently about the home when “Niko” (TM) was there and wanted to remember that period.

3.3.3.4. Hints of Personality

Householders attributed some personality to the TM. TM’s computer at household A crashed at one point during the deployment and the father was around as we rebooted the machine and restarted TM software. Even though TM had no data from household activity, TM had adapted to the household by clustering the space of possible image styles around the only data it had available – the system clock and the activity in the special zone. When TM crashed, all of these cluster centers were lost, and TM started again with randomized cluster centers. The new images were strikingly different from what the family had been seeing the last few weeks. Andrew described it, “Wow those are the old ones [dark green, blue, purple, maroon]?! ... That's what it looked like when we started, like an infant. ... [it had] more shapes inside other shapes.” (Andrew). He interpreted the simple compositions as being childlike and simplistic, while in contrast the images he had been seeing for the past weeks had been more complex and delicate. Andrew interpreted this as TM growing up (and the crash had reverted it to childhood).

Household C also provided evidence of ascribing personality and aliveness to the TM. At one point, ten-year-old Charlie had his mom buy him an embossing labeler, which he used to label people and animals in the house. Carla, Charlie, and Charlie's friend, who was over for the afternoon, were all labeled with their first name on their forehead. Charlie then labeled the two cats, and, interestingly, labeled TM screen with "Niko" (Niko is the brand of flat panel TV we provided with the TM). He did not provide a name label for any other technology artifact in the house, including the main TV, his videogame consoles, or any of the computers in the house (he did eventually label his personal laptop). The family referred to TM as "Niko" from then on, and came to treat it as a social presence living in their house (as described above).

3.3.3.5. Printing Practices

Householders, both young and old, printed many more images from TM than we originally expected, printing images that they liked, or that they wanted to comment on, or that were "new" (i.e., they showed a previously unseen color palette, or a previously unseen shape grammar). Householders also wanted to walk us through their images when we came to interview, spontaneously leading our interviewer to TM book. One of the fundamental motivations for this printing practice was the fact that images appear only for a couple of minutes and never again. This extreme fleeting characteristic of TM's paintings prompted participants to capture the moment, similar to taking a snapshot with a camera. The physical quality of the printout on bright white heavy paper raised an aura of permanent and valuable memorabilia.

In the A house (with the broken TM), print activity was, as we have noted previously, more limited. A young householder still claimed that she enjoyed the printing

activity, stating, “I might as well just print a couple pictures. And by a couple, I mean ten thousand.” (Alice) Household A, since they engaged less with TM overall, printed less frequently as time went on. In the B and C households, printing activity was constant throughout the deployment.

The C house was particularly enamored with printing; on a couple of occasions, they printed nearly 100 images per day, and sometimes 8 or 10 images per minute. Charlie would sit at dinner and print intermittently but steadily; he was constantly watching TM from the corner of his eye for “good ones” and would print the images he liked.

3.3.3.6. Deepening of Reflection

One of our design goals is for TM to be a resource for reflection. We hoped that the constantly changing stream of imagery, correlated with household activity, would help make the invisible patterns of the household visible. We found evidence TM did become a resource for reflection, though household members did not learn to interpret stylistic features of individual images in terms of household activity, instead reading meanings into colors, the overall composition, and individual shapes.

As the families lived with the TM, they began to find ways to integrate it into their routines and rhythms. In the morning, as householders got ready for the day, many family members glanced at TM as they went about their routines, printing many images during this time. Betty described the TM’s pale colors and slow refresh rate as mirroring her morning thoughts. She said, "You think about all the things you're going to do." She contrasted this with the images in the evening times, where the colors were “vibrant,” and “happy... In the evenings we're happy, since we did a good day's work" (Betty).

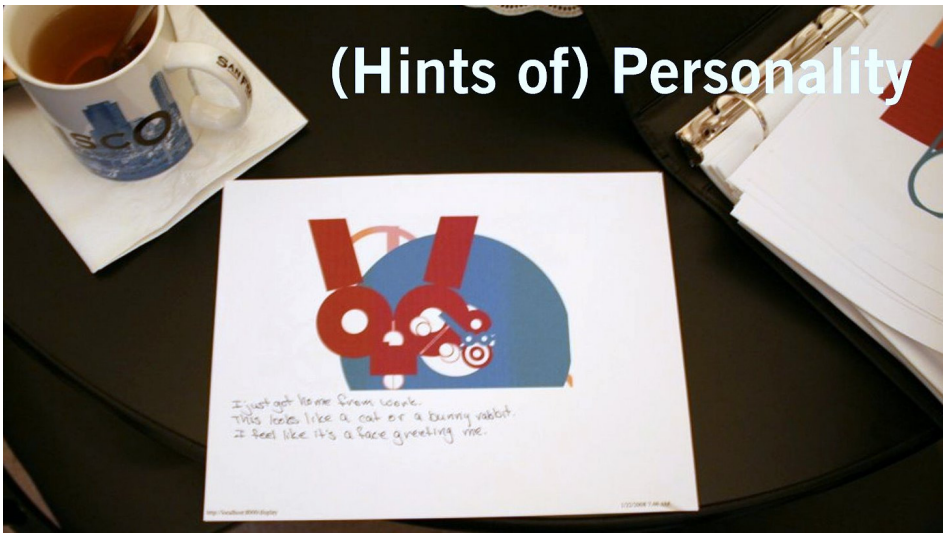


Figure 3.12: Tableau Machine's hints of personality. On the table of household C is a printout. The hand written caption from participant "Carla" states, "I just got home from work. This looks like a cat or a bunny rabbit. I feel like it's a face greeting me." The image is a production from the *OuterClust* grammar with *high coverage* and *right-heavy balance*. TM used *Rich palette 1* to color it.

Some householders reported being mesmerized for a long moment watching the TM. Alice reported watching TM instead of television, as a kind of "show" analogous to a television program. Betty noted a time when she was home alone and "trying to make myself go up [stairs]" but wound up watching TM refresh a few times first. These moments of doing nothing, just "puttering" around the house (Wyche, Taylor et al. 2007), were moments where TM became a salient resource for the unstated and even unconscious reflection on home life.

Toward the end of the deployment, Carla came home one day after her workday ended (in the early morning, so she was very tired). TM produced an image that looked like a "bunny rabbit" to her, and the bunny was looking out at her. She found this comforting and cute, and wrote on the image "I just got home from work. This looks like a cat or a bunny rabbit. We feel like it's a face greeting me" (Carla). Figure 3.12 shows the image and the text written by the participant.

In household B, a particular image appeared late in the deployment that was part of a happy moment at home. Late one evening, Byron was cooking in the kitchen and Betty was keeping him company. She walked out into the living room and an image appeared on TM that looked like a smiling face. She laughed and made of big deal of it at the time, and wrote on the image "Face of my husband. In the kitchen. Cooking. Apron." (Betty). Household B ultimately kept this image hanging from the fridge's door even after the study ended. Figure 3.11 shows the image on the fridge.

However, people did not always find relationships between images and activity. During an interview in household B, we asked about images that had been printed during an afternoon when cousins had come over to play videogames. When asked by the interviewer "Does this [image] look like video game playing?" householders, both those involved and those watching, said "No." Similarly, in household C, Carla recounted a recent evening where Charlie had done poorly on a math test, and Carla spent the evening helping him correct his work. It seemed like a tense time, so the interviewer found an image of that time and asked, "Does this look like homework?" Charlie replied that it did not.

While TM partially succeeded in becoming a resource for reflection, households did not create vocabularies around TM for describing the dynamics of everyday life. The "homeworkyness" of an evening or the "videogameness" of an afternoon remained opaque. While individual TM images would occasionally open up a moment in the home for deeper reflection, the families did not develop systematic social methods for doing so. There were interesting hints that TM images themselves became a proto-language for describing everyday life dynamics. During several interviews, householders, when

describing a particular episode in the home, would flip through their printouts of TM images to describe a household moment as "being like this [pointing at an image]."

3.3.3.7. Feelings of "Being Watched"

In all three homes, some participants, at some points in time, felt watched. We expected to find these feelings amongst householders, and were surprised that they were mentioned so infrequently. Most of the time, the TM's cameras went unnoticed. We were very careful to explain how the camera sensing of TM worked at the outset. On the pre-installation interview and at the day of deployment, we stressed to householders that the system did not store raw images. TM analyzes the images and immediately discards them.

The conditions under which participants did report feelings of being watched were interesting. In all households, adult women reported feelings of unease in some situations. Mothers in houses A and B recounted feeling as if the system was watching them as they walked around their public areas late at night. Betty specifically mentioned her pajamas and whether her body position would expose her to the cameras. Amanda mentioned eating a late night ice cream snack in the kitchen and feeling like the system "might tell on" her.

In the C household, feelings of being watched were less pronounced. Carla reported "making sure" she was dressed appropriately before coming downstairs in the mornings to make coffee. Charlie also mentioned that he called the overhead cameras "the spies," stating:

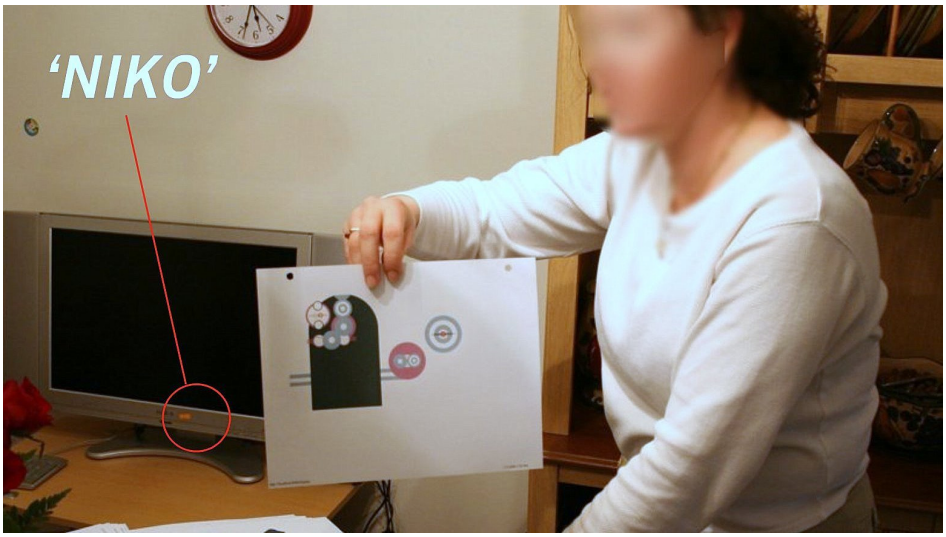


Figure 3.13: Tableau Machine (TM) in household C received the name "Niko" from the television's brand. "Charlie" placed a sticker labeling the machine and called TM, "Niko and his spies." The overhead cameras were the spies.

Charlie: [pointing up] "the spies"

Carla: "Yeah, he calls it 'Niko and his spies.'"

Interviewer: "Really, 'Niko and his spies?'" (See Figure 3.13).

Carla: "Yeah."

Next, we summarize the implications for smart technologies in the home.

3.4. Contributions: Implications for Smart Home Design

Returning to our goals of co-interpretation, we strove to create a curious and vague artifact that provides a novel and engaging window into everyday home life, creates a sense of social presence (personality), is engaging over a long period, and becomes a resource for conversation and contemplation on the rhythms and routines of the home.

For households B and C, TM succeeded in being engaging over the entire period of the study. This is markedly different from household A, whose engagement, printing

activity, and interest waned. Our inadvertent miss-configuration of TM for household A allows us to compare purely projective interpretation (the entire meaning found in TM comes from the family) and co-interpretation, where TM actively participates in meaning making. Household A's failure to incorporate TM into long-term family life provides powerful evidence that the success of TM is not purely a function of humans being able to read meaning into almost anything (a Rorschach effect), but rather that TM's active interpretation and generation supported human meaning making.

Even in successful TM households, families had some trouble describing the mood or character of their homes – the very focus of TM. While there was evidence of TM providing a resource for reflection (described above), their descriptions of activities, events, and rituals around the home were primarily factual reporting. It may be that the "Fine Art" nature of TM made it difficult for families to bootstrap a language for talking about the home; our families also had difficulty verbally describing TM compositions. They did not readily come up with design-focused descriptions of TM images such as "balanced/unbalanced," "delicate/bold," "sparse," or "juxtaposed." They used words that are more common, for example "vibrant" and "empty/full." This may have prevented them from remembering or even consciously noting some of the distinctions in TM's image space. Remember that TM maps distinctions in home activity into distinctions in the image space. It would be interesting to deploy TM in a household that includes artists, designers, or Art historians to see if this results in TM becoming a deeper resource for reflection.

We also gathered information about ways in which householders experimented with TM. Children waved at the cameras and jumped around, while adults performed

more structure experiments. One limitation is that householders experimented on the timescale of a few seconds or at most a few minutes. However, since TM aggregates motion data, the machine only subtly notices and reports the eventful experiments. Furthermore, the learning rate of TM has a cycle time of several days. While the process of figuring out TM was a long-term activity, it took place in very short bursts of reflection and experimentation. TM did not support these experiments by immediately noticing and responding to householders. In retrospect, one aspect for clear improvement for TM is to include more and lower-level interpretative scaffoldings. We overshot our target. Examples of lower interpretative scaffoldings include elements or areas on the production that directly map to regions or people in the event, elements on the screen that move like people do (left to left, forward to up), and colors in the composition that resemble the colors of people clothes. These are three examples of direct mappings that would make TM much easier to read and engage. Our participants in particular were rarely in search of hard interpretative challenges. That was not a common practice in their home culture. As we have stated, the target audience for Tableau Machine should have been High Artists.

Next, we report the larger and more broadly applicable design lessons we distilled from these studies.

3.4.1. *Activity Characterization*

With Tableau Machine, we introduce *Activity Characterization*, a practical alternative to Activity Recognition for context aware computing applications responsive to high-level, abstract activities. Traditionally, activity recognition seeks to classify activity into concrete and precise categories, starting with the classification of low-level

motion; high-level activities are recognized by aggregating lower-level classifications. In contrast, Activity Characterization directly seeks to characterize high-level activities using larger, more abstract categories that tolerate ambiguity. By bypassing the need to recognize with precision low-level activities, Activity Characterization provides a robust sensory interpretation framework for applications responsive to high-level activities. Additionally, Activity Characterization causes us to recast the activity recognition problem, opening up a new taxonomic framework for thinking about recognition. Sensing features that characterize an activity, instead performing categorical activity recognition, allows for ambiguous but still useful measurement. For problems where activity recognition may not be required, Activity Characterization provides a tractable alternative.

Activity Characterization is highly abstracted classification of activity. One way of distinguishing Activity Characterization from activity recognition is to examine the types of labels the two approaches assign to data. In the case of activity recognition, labels take the form of concrete activity names that are usually applied to data in a mutually exclusive manner. The activity labels make ontological commitments to the fundamental human activities in the domain. In contrast, the labels in Activity Characterization correspond to abstract features of the data, features designed to capture useful (for a specific application) characteristics of the high-level context. No commitment is made to intermediate, lower-level activity labels in computing the abstract features. The feature labels are not mutually exclusive. An activity instance may have multiple and overlapping characterizations. Activity Characterization would not satisfy an application that requires mutually exclusive and concrete activity categories.

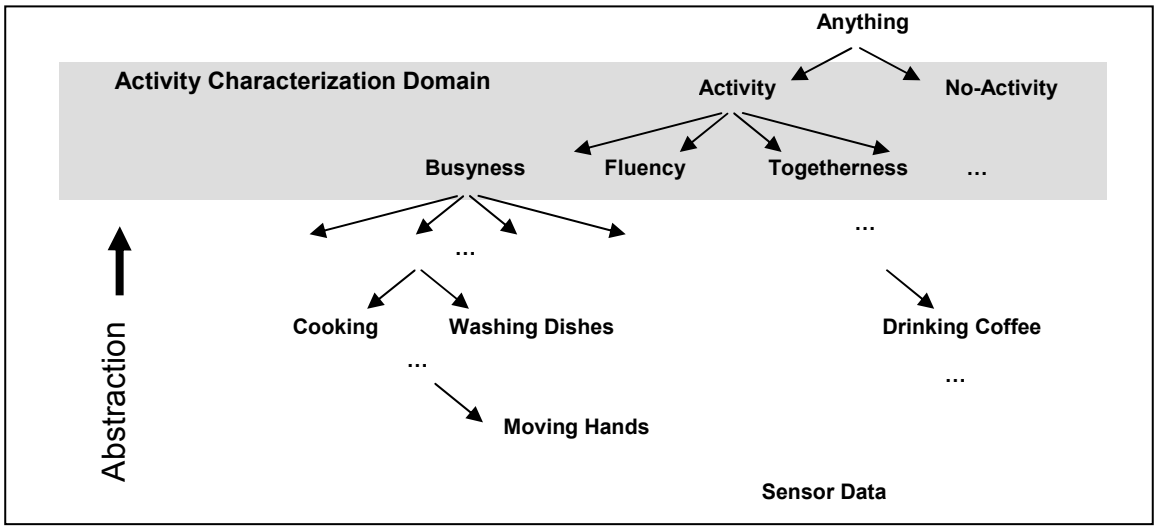


Figure 3.14: Activity Hierarchy from concrete to abstract (Activity Characterization).

A further distinction between Activity Characterization and activity recognition is the level of abstraction, still related to ambiguity. Consider the activity abstraction hierarchy in Figure 3.14, where the leaves are atomic actions into which raw sensory data classifies and with activities becoming increasingly compound and abstract as you move up the tree (review section 2.2 on page 15). Activity recognition traditionally classifies activity into the categories found near the bottom and middle of Figure 3.14. Activity Characterization, on the other hand, assigns labels near the top of the tree. The goal for Activity Characterization is to push towards the top of abstraction while remaining relevant to a particular application domain. Some applications and domains may permit higher levels of abstraction than others may.

Note that the arcs in activity hierarchy represent composition; activities that are more abstract are composed from lower-level activities. This composition, however, is not simply determinative. As one moves up the abstraction hierarchy, the activities encompass descriptions of context that are more abstract. If the compositional structure of lower-level activities fully determined the higher-level activity, that would be

equivalent to saying that context, in its full generality, is fully compositional. As a thought experiment, consider an activity such as "drinking coffee." Though it may play a role in a highly abstract activity such as "being busy" or "being together," it is very difficult to imagine how "busyness" or "togetherness" could actually be defined by grammar-like compositions of activities such as "drinking coffee." As we move up the abstraction tree, we describe increasingly encompassing characteristics of the context. For context characteristics near the top of the tree, it may not be possible, even in principle, to compute these characteristics compositionally from lower-level activities. This is why Activity Characterization computes abstract activity labels without commitment to lower-level activity ontology.

The design cost for performing highly abstract context labeling without lower-level labels is ambiguity. Ambiguity occurs when it is impossible to distinguish between multiple activity-labels at the same level of abstraction. When one directly labels activity in terms of more abstract activity classifications, the less abstract activity classifications are aliased (masqueraded or muddled); we cannot distinguish between the less abstract classes solely based on the more abstract one. Since Activity Characterization directly labels activities using highly abstract context labels, this will produce a large amount of system ambiguity about the more concrete activity categories. Such ambiguity, however, is a valuable design resource for the appropriate applications (Aoki and Woodruff 2005; Gaver, Sengers et al. 2007). It allows, for example, the partial sharing of private data. The implementation win for using Activity Characterization is that one avoids the well-known robustness and accuracy issues associated with activity recognition, allowing designers to focus on building working, end-to-end context-aware systems.

3.4.2. *Cameras in the Home*

We were very explicit in our discussion with householders regarding how TM uses camera data. In our deployments of TM, households were able to accept and even forget about the cameras for most social situations. Those situations in which cameras raised red flags imply two important facts for design teams installing sensing technologies in the home. First, different householders will have different reactions to invasive sensing; in our study, women were more sensitive than men. Second, alone time, even in social places in the home, is more sensitive than social time. Ubicomp researchers may want to use the system's sensors themselves to change the recording based on which householder is in the sensor's view and whether a situation is social or individual.

3.4.3. *Mental Models and Experimentation*

Ambient intelligent systems should support rapid experimentation by household members. We found that families actively experimented with TM to more deeply understand the system. However, these experiments were only a few seconds to a minute long; the longer time scale on which TM responds to activity made it difficult for families to perform successful experiments. Ambient intelligent systems should have interpretive scaffolding modes that support active experimentation by responding to short-term activity. They should come closer to the paradigm of direct manipulation when people are in experimental mode.

3.4.4. *Enhancing Experiences with Co-Interpretation*

Users are naturally curious and playful. This curiosity extends beyond the first few hours or days and can be extended through careful design. Users' interactions with a

pervasive system (and feelings about it) will change over a long-term deployment. Paying attention to the playful and experiential aspects of a system can help it to become a fixture in the home. Co-interpretation need not be restricted to playful and artistic systems, but could be used to enhance task-based systems as well. For example, a cooking support system could help users experience differences in the felt-life of cooking (hurried vs. leisurely, social vs. alone) while also providing task support.

3.4.5. *Printing as System Feature and Evaluation Aid*

The ability to print system states was successful feature in TM. It worked both as an engaging activity for householders and as a way to evaluate and analyze householder reactions. Even in task-based software, printing of system state can be great way to understand what is not understood by users, as well as to get rich accounts of what they were trying to do at that moment. The prints served as a memory aid to reconstruct the situation, as well as a souvenir. Participants were more than willing to write on the printouts, denote important or strange parts, and describe their intentions and questions.

3.5. Conclusions

In this chapter we have presented the evolution of Tableau Machine's design through formative evaluations and the summative and longitudinal evaluation of Tableau Machine. We created Tableau Machine to answer the fourth research question associated with the fourth claim of our thesis statement. Can a vision-based visualizing Art installation engage users in a long-term process of creative interpretation, experimentation, conversation, and reflection? Succinctly, Tableau Machine is an interactive Art installation that perceives and interprets its environment and generates

novel paintings based on its interpretations. The primary goal of Tableau Machine is to engage its co-occupants in long-term creative interpretation.

In this chapter, we discussed the qualitative findings of three installations, two of which worked properly. The installation with miscommunication between sensing and generating afforded an impromptu control point to compare the other two installations against. Our main findings are that people welcomed only the working versions of Tableau Machine as a pet into their homes. They created entertaining stories around Tableau Machine's output and assigned it intentionality, mood, and personality. Participants experimented systematically and enduringly with the input-output mappings of Tableau Machine, creating partial models of its workings. In some instances, participants decoded two of the three increasingly complex mapping levels. Nevertheless, the full depth of Tableau Machine's mappings remained beyond all the participants.

Finally, we analyzed a number of contributions to smart home design. For example, Tableau Machine opens a wider design space for the sensing of human context, showing that valuable interactions can be produced from ambiguous and imprecise metrics. The most valuable contribution of Tableau Machine for this thesis is the creation and utilization of the Activity Table. Visualizing aggregate motion across place and period fuelled the designer's creativity. The table opened deep and useful insight into people's everyday behaviors. The natural questions were, "could it work for other users and tasks?" "What other views and functionalities could it afford?" "What challenges would other users face?" After Tableau Machine, we steered this research toward pure visualization of activity through computer vision: Viz-A-Vis. In the next three chapters we describe in detail the design cycles and three evaluations of Viz-A-Vis.

CHAPTER 4

VIZ-A-VIS: SUPPORTING HUMAN ANALYSIS OF ACTIVITY IN NATURAL SETTINGS OVER VARIABLE PERIODS OF TIME

In the established procedural model of information visualization, the first operation is to transform raw data into data tables (see Figure 2.1). The data transforms typically include abstraction steps that aggregate and segment relevant data. In this model, the human operator usually defines the transforms. The theme of this chapter is that for raw video data, data transforms may be supported by robust and low-level computer vision. The high-level reasoning still resides in the human analyst, while the low-level perception is handled robustly by the computer. To illustrate this approach, we present Viz-A-Vis (**VIZ**ualization **A**ctivity through **VIS**ion), an overhead video capture and access system for behavioral and occupancy analysis in natural settings over long periods (Romero, Summet et al. 2008). Overhead video provides a rich opportunity for long-term behavioral and occupancy analysis, but it poses several considerable challenges. First, it generates very large volumes of video data, which are impractical to analyze manually. Second, automatic abstraction of high-level semantics from overhead video remains an open problem for computer vision, machine learning, and pattern recognition. Third, cameras are intrusive sensing technologies. In this dissertation, we present the initial steps addressing these challenges.

4.1. Goal

We present Viz-A-Vis, or **VIZ**ualization of **A**ctivity through **VIS**ion (see Figure 4.1), a capture and access system (Truong, Abowd et al. 2001) that serves as an initial approach to building information visualization interfaces on top of computer vision abstractions to take advantage of the opportunities and tackle some of the challenges of continuous overhead video analysis. Our focus is on bridging the semantic gap between insightful high-level human analysis and robust low-level machine sensing (Hare, Lewis et al. 2006) through a mixed-initiative computing approach. Mixed-initiative computing systems are symbiotic human-computer systems that take advantage of the strengths and limit the weaknesses of its constituent actors. From the machine's side, we bridge the semantic gap through computational perception methods. From the human's side, we bridge the semantic gap with information visualization methods. Bridging the semantic gap with machine vision alone has remained an open problem for decades. Bridging the semantic gap with visualization alone requires significant manual work from the analyst and is an impractical solution for analyzing long-term video. Through Viz-A-Vis, we explicitly tackle the first two challenges. For the third challenge we assume application proportionality (Iachello and Abowd 2005), that is, the perceived benefits of select applications outweigh the cost of lost privacy.

The overarching goal of our work is to provide an iterative infrastructure that improves this analytical process from three perspectives. First, we gain greater insight into overhead video data and its utility. Second, we gain experience in the design of the information visualization interface to this type of data. Third, we improve the computer vision design process by providing a high-level visualization of low-level data and

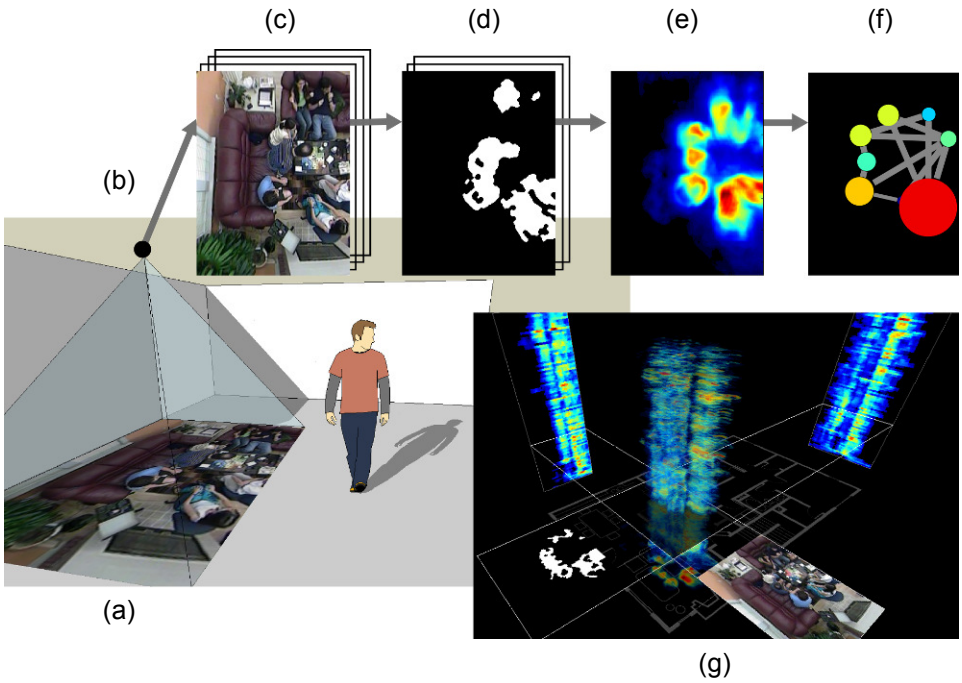


Figure 4.1: VIsual**i**Z**ation of **A**ctivity through **VIS**ion (Viz-A-Vis) overview: (a) place of interest; (b) overhead camera; (c) image sequence; (d) motion sequence; (e) Activity Map, spatial and temporal aggregation of motion; (f) semantic aggregation of motion; (g) Activity Cube, visualization of aggregate motion over space, place, and time.

algorithm workings that may be used as better feature vectors and target patterns for machine learning and pattern recognition.

In the following section we briefly explore the theoretical foundations of our method for bridging the semantic gap. In section 4.3 we describe in detail the architecture of Viz-A-Vis, from capture to analytical insight. In section 4.4 we present a preliminary case studies where Viz-A-Vis opened new insight into behavioral patterns. Finally, we conclude and present contributions.

4.2. System Architecture

Viz-A-Vis is a capture and access system, where the capture comes from overhead cameras and the access is the analytical process mediated by information

visualization on top of computer vision (see Figure 4.1). The raw data captured goes through two inverse processes: a process of abstraction, where relevant data is automatically segmented and aggregated, and a process of reification, where visual overviews are explored, filtered, zoomed, contextualized, annotated and indexed back to relevant video sequences. The goal is to provide a visual roadmap that serves as a video semantic navigation tool.

4.2.1. *Process of Automatic Abstraction*

From a theoretical perspective, the raw data for sensing infrastructures is the real world. Thus, the process of abstraction begins at the selection and placement of sensors. There are usually many competing considerations, such as expressiveness, relevancy, and intrusiveness. If the sensor does not have enough expressive power to capture all target events or if the target event does not exhibit enough observable phenomena in the modality of the sensor, recall from the real world will never be complete. Although this is not a central topic for information visualization, it is an important consideration to keep in mind, especially when realizing that the insight we are looking for is about the real world, and not the mediating sensor data.

For Viz-A-Vis, we begin the process of abstraction with the areas of coverage (Figure 4.1-a). We have installed the system in a research laboratory, four area homes, two living laboratories, and two museums. In each installation we carefully analyzed the space, the objects in it, and the occupancy of the space, through preliminary interviews.

We chose cameras because of their expressive power. Our assumption is that all visually observable human behavior, down to single fingers moving, can be captured by a camera. We chose to place the cameras over the areas of interest for several practical

physical and algorithmic considerations. Physically, by being in the ceiling, the cameras are relatively out of sight and out of the way. Algorithmically, by having an overhead view of the world, the computation of low level vision percepts is greatly simplified.

In video, the process of abstraction begins at the hardware level, with quantization and discretization of time (frame rate), space (resolution), luminance (sensitivity to light), and chrominance (sensitivity to color). The camera should have the speed, resolution, and sensitivity to capture all target behaviors in its field of view for its intended application (Figure 4.1-b). The choice of camera is an important consideration when applying Viz-A-Vis. For our applications, we used off-the-shelf cameras and ran them at relatively low frame rates, between 1 and 1.5 frames per second, relatively low resolutions, between 160 x 120 and 640 x 480 pixels, and normal 24 bit color (Figure 4.1-c). We changed the lens to a 120° field of view, wide angle lens to increase coverage.

The next abstraction step is the computer vision we apply to the raw video, the central theme of this paper. Of the innumerable techniques available in the vision literature, we purposefully chose to restrict our abstraction to motion (Figure 4.1-d). Motion is considered one of the most robust and lowest level abstractions from video. Furthermore, overhead video readily affords a number of important technical simplifications to the computation of motion. First, the camera is fixed, both in its internal and external parameters (focal length, position, and orientation). Second, the frustum is vertical.

These two simplifications mean that we can, in practicality, assume there is a one-to-one correspondence between image and architectural space and that there is a single plane of interest, the ground. Ignoring the error introduced by parallax, mapping pixels to

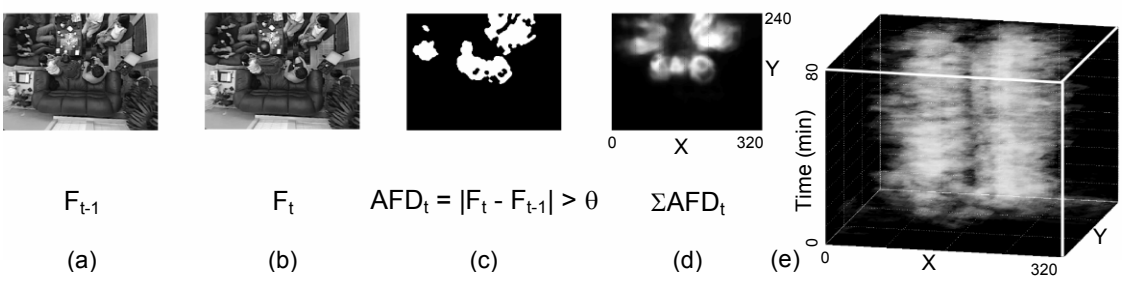


Figure 4.2: Computing and aggregating motion by adjacent frame difference (AFD): (a) previous frame; (b) present frame; (c) adjacent frame difference (AFD); (d) sum of AFD over time; (e) Activity Cube, partial aggregate motion layers across time.

small areas in physical space is a simple, realistic, and robust abstraction. Third, in natural settings, changes in architectural space (image background) are rare events. Fourth, dramatic illumination changes occur very sporadically. Fifth, the likelihood of people appearing identical to the background is extremely low. At least some part of their body will be of a different color, shade or texture than the background. And sixth, the likelihood of people holding perfectly still drops to zero very quickly. Under these practical conditions, we compute motion from video by simple adjacent frame difference (AFD) and we associate this motion with the physical space it occupies.

We subtract gray-scaled adjacent frames in time (Figures 4.2-a..b) and threshold the difference (Figure 4.2-c). The result is a binary motion image, where white pixels represent motion. We clean up the binary image with the morphological operators open and close. The threshold and the morphological operators serve as signal-to-noise ratio control parameters.

The binary motion image is much smaller than the original frame, yet it contains most of its semantic relevancy. It shows when, where, and how much motion occurred. As a concrete example, consider a 640 x 480-pixel, 24-bit frame. It contains 7,372,800 bits. A binary motion image of the same resolution contains 307,200 bits. Typically,

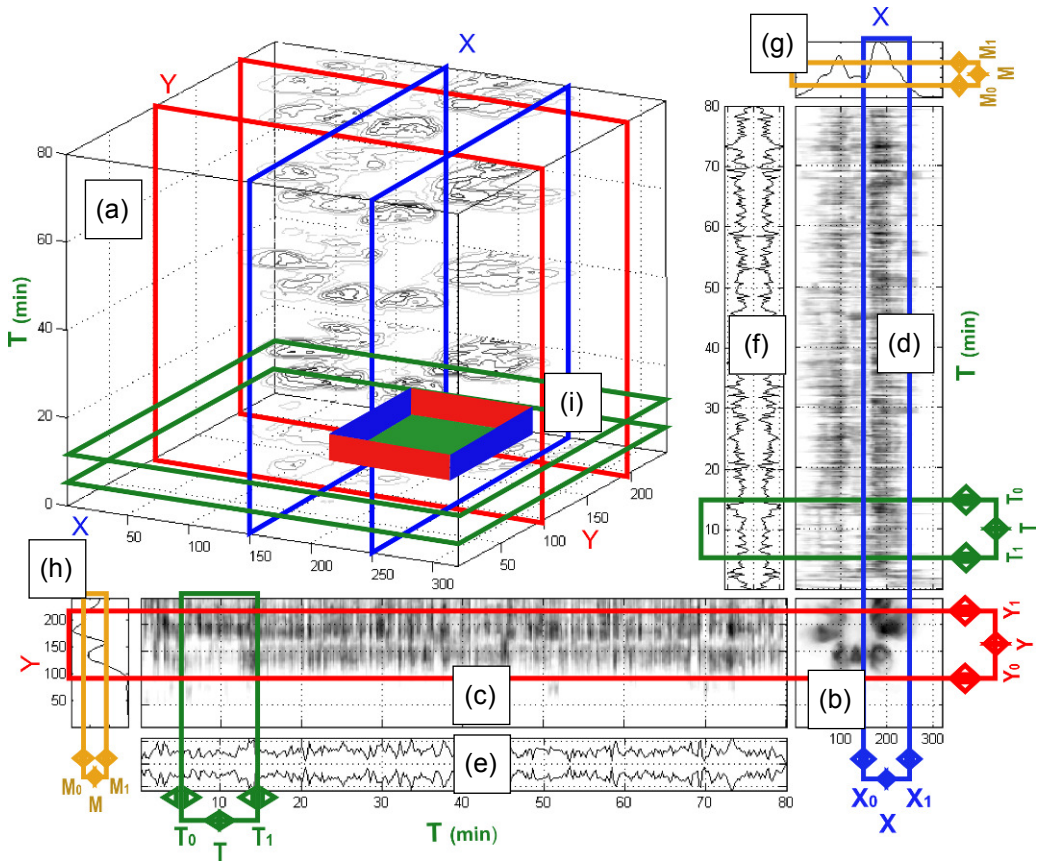


Figure 4.3: Model of visualization and navigation for the Activity Cube: (a) Activity Cube showing 5 aggregate 2D isocontour slices of motion across 80 minutes; (b) Activity Map, aggregation of motion across entire 80 minutes; (c) aggregation of motion across X (Y vs. T); (d) aggregation of motion across Y (X vs. T); (e-f) aggregation of motion across X and Y; (g) aggregation of motion across Y and T; (h) aggregation of motion across X and T; (i) sub-space result of the query (X_0<X<X_1) & (Y_0<Y<Y_1) & (T_0<T<T_1). The dynamic query is performed through double sided sliders on X (blue), Y (red), and T (green). The fourth querying dimension is aggregate motion M (yellow).

these motion images are sparse. Assuming 5% percent of the pixels are active, a typical motion image can be encoded in roughly 138,240 bits. This is an abstraction that hides roughly 98% of mostly irrelevant data.

Since image space has a one-to-one correspondence with physical space, we can easily aggregate the data over space and time (Figures 4.1-f..g, 4.2-d, and 4.3-b..h) and we can stack the motion frames so that time is represented in the third axis of a motion cube people across image space, physical space, and architectural space across time. The

Activity Cube and the aggregates we compute from it serve as the basis for our visualization (Figure 4.1-g).

Figure 4.3 shows Viz-A-Vis' model of visualization and interaction with the Activity Cube. As with other 3D visualizations, the cube presents a number of challenges. Because of perspective and occlusion, to get a clear picture of the structures, we need to be able to rotate, translate and zoom the view in three dimensions.

We use the cube as a high level overview to the data and provide a number of marginal aggregations that serve as 2D and 1D "x-rays" of the cube (Figures 4.3-b..h). These aggregates are higher abstractions of the data. Next, we augment these aggregate views with dynamic querying capabilities through double sided sliders. Finding target events in the cube is equivalent to defining the relevant spatial and temporal boundaries of a sub-space or manifold inside the cube (Figure 4.3-i). At this stage, the only possible shape of the sub-space is an orthogonal parallelepiped. In reality, finding target events may require following translating motion across space. These types of events would be snake-like 3D manifolds inside the cube. Simple orthogonal query sliders are unable to capture such structures. To coarsely achieve this, a first approach is to augment the conjunctional queries with disjunction capabilities.

So far, we have presented purely spatial and temporal abstractions. These abstractions segment relevant semantics, but are not intrinsically semantic. The final level of abstraction we present in this paper is aggregation over places of interest. We define places (or regions) of interest manually. They could be defined dynamically and automatically, but we wanted to keep control of this process with the human at this first stage. We segment image/physical space into meaningful regions. We start with the

observation that place is socially meaningful space. Our first method is to divide the image space into architectural elements of the space, such as hallways, doorframes, chimneys, kitchen counters, and appliances. This is equivalent to segmenting the Activity Cube into pre-defined orthogonal parallelepipeds spanning the height of the cube. Next, we divide the space based on large furniture such as the couch, the coffee table, the dining table. We call these divisions semantic activity zones (SAZ) (Romero, Pousman et al. 2007). In all our observations, these definitions remained stable throughout the deployments, even up to 6 months. If the furniture layout changes, though, there are simple computer vision algorithms to detect and track those changes. The furniture has fixed appearance since its distance to the camera remains relatively constant and there are no out of plain rotations. We did not address this automatic tracking since our deployments did not require it.

In Figure 4.4 shows another version of the Activity Table in Figure 3.4 on page 37. Both contain the same data. This version of the Activity Table maps the aggregate of motion over places of interest across time onto the intensity level of its rows across its columns, respectively. More generally, the Activity Table is a tabular representation of semantically aggregated motion across time. Figure 4.4 shows the floor plan of the Aware Home on the left, Georgia Tech's living laboratory (Kidd, Orr et al. 1999). Figure 4.4 also shows the manual segmentation of the floor plan into SAZs. In this space we defined 39 zones. To highlight a couple of interesting examples, zone 15 is the living room sofa in front of the television that is mounted above the fireplace (zone 13). Zone 23 is the dining room table. The Activity Table shows the activity of the 39 SAZs labeled on the left. The image streams come from 10 cameras, 4 in the living room, 2 in the

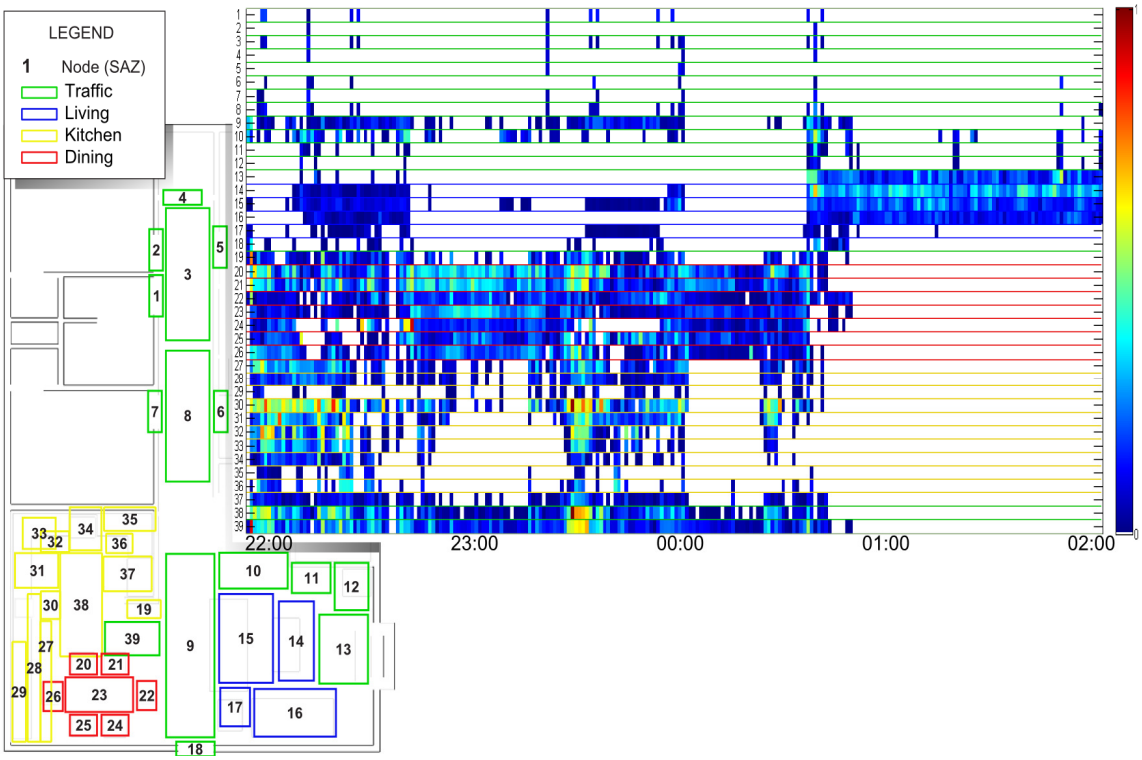


Figure 4.4: Floor plan, semantic activity zones, and Activity Table.

dining room, 2 in the kitchen, and 2 in the hallway. Figure 5.11 shows the frustums of the ten cameras. We color coded the zones based on the regions they belong to: kitchen is yellow, dining room, red, living room, blue, and transit green. We added the color-coding to the rows of the Activity Table.

Note that the adjacency relationship between zones in the floor plan is two-dimensional. By aligning the zones along a single column, some adjacency relationships are lost. For example, zone 9 is adjacent to 8, 10, 15, 18, 19, 22, and 39. In the table, it is adjacent to 8 and 10 only. Thus, in order to visually track changes in location it is necessary to skip rows. This can be mitigated by row re-ordering or hiding. The problem with reordering and hiding is that part of the process of learning to read activity in this table relies heavily on row stability.

The data shown on this instance of the Activity Table is a dinner party of eight adults. They prepared dinner, ate, cleaned up, and played a game board in the living room. The data that we have shown in Figures 4.1, 2, 3, 4, and 6 come from the lower right camera in the living room and from the period where the 8 adults played cranium.

The Activity Table is highly abstracted. It allows us to visualize five hours of data coming from ten cameras at 1.5 fps and 320 x 240 resolution in a single 2D view. Without abstraction and excluding color, there are 768,000 variables. With this abstraction, there are 39. We have eliminated 99.995% of the complexity. Of course, this reduction comes with a price.

The Activity Table is an effective visualization for large motions across space. The transitions between kitchen, dining room, and space are very apparent. We label this type of motion translation. The Activity Table, on the other hand, is not as an effective visualization for motion that does not produce a change in location. We call motion that occurs over the same space vibration. It is hard to distinguish fine events inside the large episodes annotated in Figure 3.4-a. For example, during the game of Cranium™, there is a finer granularity that is lost in this visualization. The game has turn taking, it has different modalities of play, and it has different outcomes at each turn. All of these behaviors are washed out at this level of abstraction.

We experimented with several techniques to avoid losing sight of vibrations, including zooming and finer granularity for the parsing of space, a type of semantic zooming. These techniques help, but are not enough. We now present the process of reification, the practice of going from abstract to concrete representations.

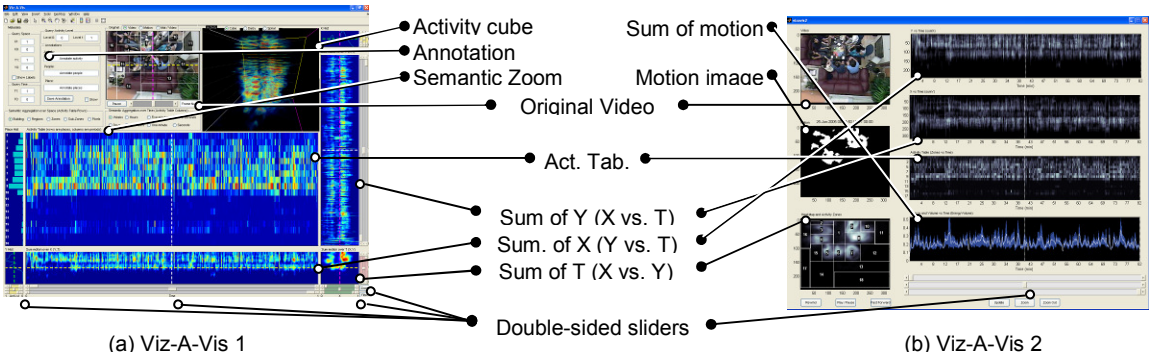


Figure 4.5: Viz-A-Vis formative evaluation prototypes: (a) prototype 1 and (b) 2.

4.2.2. *Process of Interactive Reification*

Up to this point, the only input from the analyst is the definition of semantic activity zones. We now describe in detail the types of exploratory interactions we designed for Viz-A-Vis, which serves as a reification toward the relevant raw data. At the abstract level users make hypothesis that they reify and test by looking at the original video.

Figure 4.6 shows the final interface for Viz-A-Vis. It is a geographical information system (GIS) where the geography is the floor plan of the environment, annotated with simple outlines of the furniture and spaces contained within it. The layers stacked on top of the floor plan are aggregate slices of motion across time. The data in Figures 4.5 and 4.6 come from the bottom right camera in the living room during the episode of playing cranium at the end of the events in Figures 3.4 and 4.4.

This GIS-style visualization is the third prototype of a sequence we formatively evaluated through interviews with 8 information visualization researchers. We presented the three prototypes to each expert, explained the data, the analytical goals, the transformations and the views. The first prototype unfolded the orthographic aggregates horizontally and vertically (see Figure 4.5-a) and downplayed the view of the cube in preference of the Activity Table. All but one of the reviewers found integrating the

vertical and horizontal views of time awkward. The second prototype showed all aggregates across time horizontally, from left to right. The downside of this is that the X vs. T aggregate view is transposed and maps left to up and right to down (Figure 4.5-b). Integrating the spatial information continued to be a challenge. We arrived at our GIS visualization for two main reasons: first, the visual integration of the aggregate views is simpler under 3D perspective; second, the floor plan provides valuable context for visually disambiguating the Activity Cube and its aggregates.

We will now review the design of the third prototype. First, we provide high level overviews in the Activity Table on the left and the Activity Cube. The Activity Table is not part of the 3D structure and sits in front of the cube. Rotations and translations do not affect the table. The user can brush space, place, and time on both views, though, and zooming and filtering on either will affect the other and all the other views of the orthogonal aggregations. The Activity Table on the left of Figure 4.6 is a transpose of the table in Figure 4.4. Time flows up. The SAZs are the columns of this table, and time goes from bottom to top, in the same direction of the cube. It seems more natural to show time starting at the ground and advancing up without boundaries.

Directly on top of the ground we show the Activity Map, a heat map aggregating the entire period being considered. Together with the outline of the floor plan and the furniture on it, this temporal aggregate serves as an effective summary of the activity during the time period at hand. Unfortunately, it hides the sequence of events. There are techniques that show aggregate and sequence of motion, for example, temporal templates (Bobick and Davis 1996). This technique fades the motion as time goes by.

Unfortunately, it does not scale well for long and complex sequences where multiple actors occupy the space under observation.

Separated by a prudent gap to avoid occlusions, the Activity Cube lies directly above the temporal aggregate and the architectural space it tracks. Here, we are showing the same data as in Figures 4.2-e and 4.3-a. Since the motion captured in this video sequence is vibration, the Activity Cube naturally forms cylindrical columns in the places where people sat.

We aggregate the data into roughly one-minute slices. The temporal window of aggregation is an important parameter of the visualization. Different temporal patterns will emerge at different aggregation granularities. Some patterns will emerge with a two-second aggregation window, like loading the dishwasher, while other patterns will emerge with a one-day granularity, like weekdays versus weekends (see Figure 6.19 on page 201). Furthermore, the number of temporal slices is constrained by the space and resolution of the display screen. For Viz-A-Vis we compute by default a discrete optimal aggregation window as a function of the length of the sequence and the size of the screen. We also allow the user to manually define the aggregation window if needed. We double map each heat map layer in the cube to color and opacity. Thus, areas with lower aggregate values will be simultaneously darker and more translucent. We experimented with several views, including voxel representations, isocontours, and isosurfaces. Translucent aggregate slices maintained the visual structure of the data better than the other options.

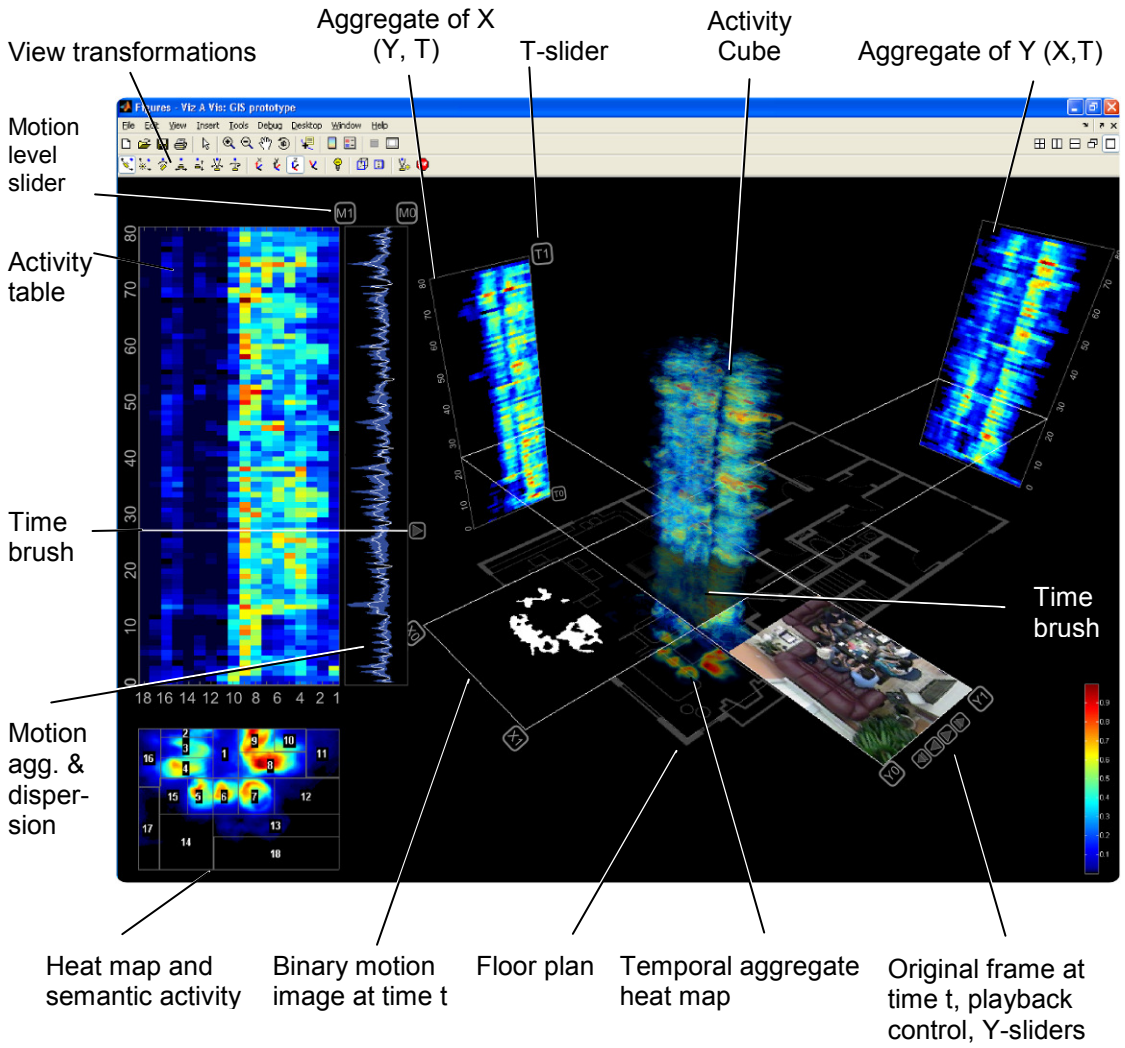


Figure 4.6: Viz-A-Vis interface. *Overview* : Activity Table, Activity Cube. *Zoom*: double-sided sliders for dynamic query on time and space. *Filter*: motion level double-sided sliders, cube translucency, and opaque time brush surface on cube. *Detail, index* and *focus*: binary motion image and original frame at time t with playback controls. *Context*: floor plan, Activity Cube, temporal and spatial aggregates. *Temporal aggregation:* heat map. *Spatial aggregation*: X vs. T and Y vs. T. *Semantic aggregation*: semantic activity zones definition and Activity Table. *Semantic Zooming*: Activity Table. *Brushing*: time brushing. *View transformations*: 3D-view rotate and translate, camera roll, pitch, yaw, position, and field of view, and variable illumination from multiple lights.

On the “walls” of the GIS we show the aggregate of motion across X and Y. They serve as x-rays of the Activity Cube. The offer navigation and contextual affordances through brushing and dynamic querying over time.

We’ve extracted the original frame and the binary motion image at the temporal point of brushing. This rapid indexing provides detail and focus and maintains the temporal and spatial context. It lets the user interpret the video data from the source. The images are laid out horizontally, as if cards drawn from a deck. The user has the option of hiding this detail. The analyst can brush the cube and pull out the original data by scrubbing with the mouse over the temporal brush. We provide typical video playback capabilities as well.

On the left hand side of Figure 4.6 are three 2D graphs: the Activity Table, the aggregate and dispersion of motion, and the Activity Map with the semantic activity zones overlaid. The heat map of activity aids the user define the regions of interest in the X-Y plane. It provides a high-level view of real usage patterns over the space of interest. Together with the floor plan, they help discover the real and dynamic social semantics of architectural space.

We conclude this section with a description of the line-and-area plot of the aggregate and dispersion of motion on Figure 4.6. The white line in the plot encodes the aggregate of motion over the entire space of observation. It is a very high-level summary of the amount of activity in the scene. The plotted blue area in the same axis encodes the dispersion of motion over the semantic Activity Table. It measures how compact or disperse the motion is. It helps differentiate similar motion aggregates resulting from different behaviors. For example, a single person moving rapidly may generate the same

motion aggregate as numerous people moving slowly. The dispersion of multiple people will be higher. We approximately compute dispersion by thresholding the Activity Table and summing the pair-wise distances between non-zeros elements. This definition and approximation to dispersion is one example of higher level semantics from computer vision and pattern recognition. Together with the motion aggregate, these abstractions have proved instrumental in the analysis of this time series.

4.3. Preliminary Case Study with Viz-A-Vis

We present a preliminary case study of applying Viz-A-Vis to understanding behavior. The study explores the effect of three different projection technologies on groups of people collaboratively interacting with a projection surface. We report our application of Viz-A-Vis to the problem of understanding the effect of three different Virtual Rear Projection technologies (Summet, Flagg et al. 2007) on a collaborative group of users working with an interactive projection surface. The goal of virtual rear projection (VRP) is to simulate the experience of true rear projection without sacrificing the physical space necessary for it. A VRP system aims to eliminate shadows on the projection surface and prevent light from falling on objects (such as users) other than the projection surface.

Figure 4.7 (top row) illustrates the three experimental conditions: Single Projector (SP), Passive Multiple Projector (PMP), and Active Multiple Projector (AMP). Single projector and passive multiple projector simply mitigate shadows on the surface by off-center projection and redundancy. Only active multiple projector corrected for shadows on the board and for light falling on other objects.

In the study, five groups of three to five people were asked to work on a collaborative task at a large interactive display for fifteen minutes, split into three five-minute sessions, one for each projection technology. We recorded overhead video for each condition, recorded camcorder video with audio for manual analysis, and collected self report data from questionnaires and interviews.

We explored the data through the different spatial, temporal, and semantic aggregations of Viz-A-Vis. The aggregate that revealed the most interesting and succinct patterns was the temporal aggregate heat map over the space in front of the projection surface. We show this heat map for each condition across the second row of Figure 4.7. The heat maps revealed trends that were not visible when watching the groups operating live in real-time, through a camcorder recording, or even through manual analysis of the raw overhead video.

In the SP condition (left column), users are clearly split by the projected light (entering diagonally from the bottom right towards the SmartBoard located at the top center) which results in the large (blue) area showing minimal activity near the middle of the room. The people to the right of the projector beam are standing forward, towards the wall and away from the projected light. The PMP and AMP conditions also show a bi-modal distribution, but those groups are much closer together, and when compared to the SP condition, the right group is not pushed as far forward. Part of the functionality of Viz-A-Vis is to be able to take individual views and extract them from the GIS. Being able to see the aggregate motion side by side, organized by condition, allowed us to notice that the AMP condition appeared to be even less split than the PMP condition.

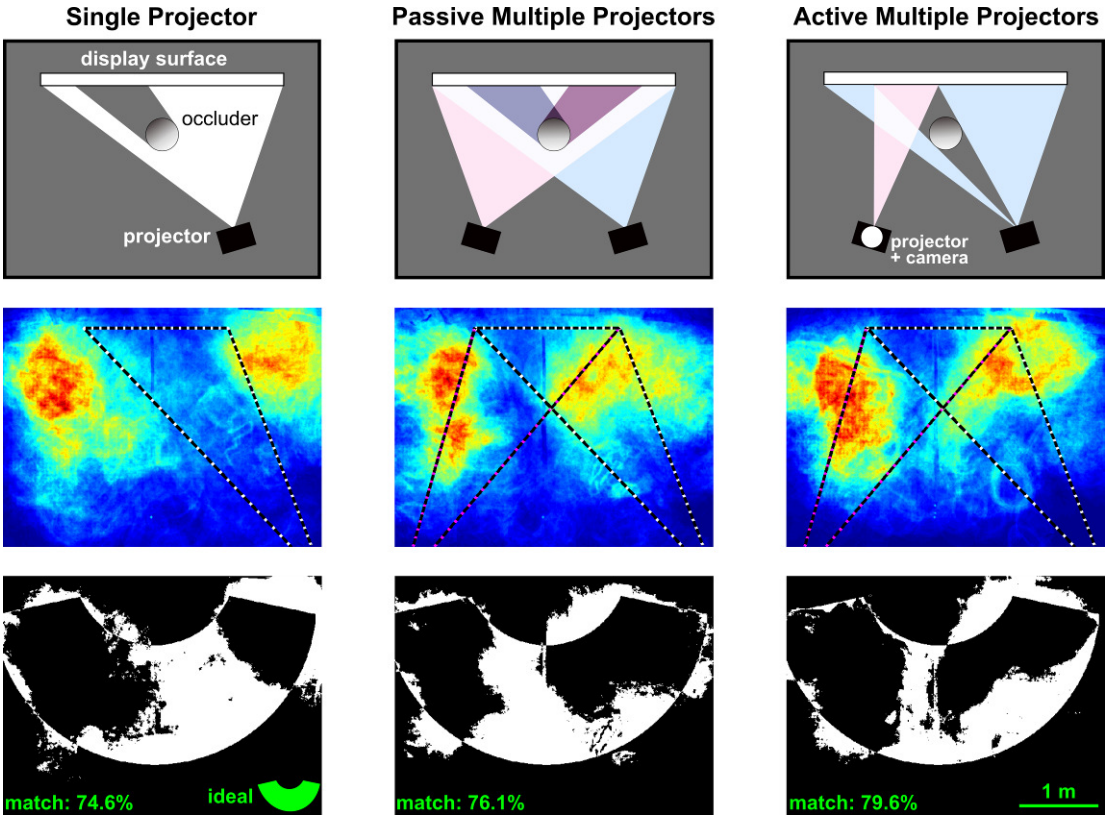


Figure 4.7: Viz-A-Vis visualizations. The three columns correspond to the three testing conditions of Virtual Rear Projection. The first row explains each technology. Row two visualizes aggregate motion. The third row visualizes template matching to "ideal" model. The percentages correspond to the match.

From this visualization we derived the concept of an "ideal" model of space usage for collaboration and used this model to quantify the space usage for numerical comparison. As we stated at the start of the paper, our third goal for the Viz-A-Vis approach is to find new features and patterns that can improve the computer vision. The ideal model we describe here is an instance of a visual pattern we discovered which can be used to advance the computational perception.

We noted that users in all three conditions where approximating a semicircular arc before the SmartBoard. We developed an "ideal" space usage model, the semicircular arc shown superimposed on the bottom row of Figure 4.7, because (1) the hole in the center

allows all users equal view and physical access to the board, and (2) the circular shape also allows equal social access to other participants. This arc is an abstraction step chosen by the analyst, a deliberate introduction of bias to gain rapid abstraction. We used a template match by sum of square differences to compare the actual study data to the semicircular arc model.

Sum of square differences is a metric of the difference between the average activity in each condition and the ideal model. This calculation is shown graphically in the bottom row of Figure 4.7. As the conditions' match-to-ideal progress from SP (74.6%) to PMP (76.1%) and AMP (79.6%), the occupancy approaches the abstract ideal. This monotonically increasing value surprised us, since the totality of user self report preference data ranked PMP well above the other conditions. The ability to aggregate user motion over time allowed us to understand how the projection conditions affected user's space usage, develop a mental model of an "ideal" space usage pattern based upon actual data, and discover that user behavior in the AMP condition matched this model closer than in the PMP condition. This analysis motivates further study of the behavioral differences between the PMP and AMP conditions. In this application domain Viz-A-Vis enhanced the analysis of previously clouded phenomena of human behavior.

4.4. Contributions

With the work we have completed with Viz-A-Vis we have opened the door for a larger research agenda. The question is: can pattern recognition technologies have a measurable impact in the visualization and understanding of voluminous raw data and can visualization techniques have an impact in the development of pattern recognition

technologies? That is a larger agenda than the one we will pursue in my dissertation. The first contribution of Viz-A-Vis is an initial step toward this research agenda.

Concretely, we have demonstrated practical methods for abstracting overhead video data in order to visualize only relevant data. We have used robust motion detection to highlight relevant semantics in the raw video data and we have transformed the problem of flat video visualization into a three-dimensional visualization that mitigates the self occlusion of the solid video cube.

We have applied the techniques of temporal aggregation of motion to the controlled experiment of three virtual rear projection technologies and we have developed a qualitative-quantitative analytical method to measure and explain the group behavioral changes dependent on the three projection technologies.

The contribution of my research is to measure the impact of computer vision technologies applied to the problem of overhead video visualization for activity analysis through detailed user evaluations. We will describe in detail the proposed empirical user study in the next chapter and the domain expert evaluation in chapter 6.

CHAPTER 5

EVALUATING THE TASK-CENTRIC IMPACT OF VIZ-A-VIS IN USER PERFORMANCE AND PREFERENCE

This chapter presents the controlled performance test of Viz-A-Vis in a task-based comparison against standard video playback and 3D video cube visualization. We compare the capacity of the three tools to accomplish five tasks: (1) describe events; (2) find the beginning of long events; (3) find short sporadic events; (4) count short motions; and (5) track individuals. The study is a within-subject evaluation. Twenty-four participants interacted with the three tools in counterbalanced order to avoid learning effects on the conditions. The objective measures are: (1) time to task completion; (2) precision; (3) recall; and (4) coverage. We collected the objective measures directly from task performance. The subjective measures are: (1) task-based ranking; (2) verbal justification for the ranking; (3) design choices for a hypothetical video forensics system; and (4) unsolicited comments, suggestions, and critiques. We collected the subjective measures 1, 2, and 3 through a questionnaire and a semi-structured interview after each participant interacted with all three conditions. We collected comments, critiques, and suggestions throughout the study.

We highlight three important findings. First, with statistical significance, Viz-A-Vis outperformed standard video playback five-to-one and the video cube two-to-one in task three, search short sporadic events. Second, both Viz-A-Vis and the video cube outperformed standard video playback two-to-one in task two, find the beginning of long events. Third, in the subjective measure 3, the hypothetical design of an airport video

forensics system, the only tool unanimously chosen for the system was the Activity Cube, the central element of Viz-A-Vis. Users cited overviewing and discovering outlier patterns as the primary tasks for the Activity Cube. The study excludes culturally or behaviorally focused observation. It omits inductive analysis such as classifying, synthesizing, or formalizing theories about activity.

Section 5.1 revisits the thesis statement and the general research questions that this study tackles. Section 5.2 describes the design of the study. Section 5.3 presents the analysis and results of the study. Section 5.4 discusses the threats to validity of the study. Section 5.5 presents the conclusions and contributions of the study.

5.1 Research Questions

Consider the overall thesis of this work:

In the process of overhead video interpretation and analysis of activity, combining computer vision abstractions with information visualization techniques provides: (1) improved user task performance measured by time to task completion, precision, recall, coverage, and user assessment; (2) improved user experience measured by user preference; (3) increased user capacity to discover activity patterns; and (4) new opportunities for creative interpretation, experimentation, conversation, and reflection regarding everyday activities. This study addresses claims 1 and 2. The broad research questions this study engages are:

Can computer vision abstractions and information visualization techniques improve the interface to analyzing activity in overhead video as measured by time to task completion, precision, recall, coverage, and user assessment?

Can computer vision abstractions and information visualization techniques improve the user experience of activity video-analysis as measured by user preference?

5.2 Design of the Study

This user study focuses on measuring the capacity of Viz-A-Vis to support human analysis of activity captured through overhead video. It is a within-subject summative usability test. In order to control order effects such as learning, we counterbalanced the three conditions of this study: (A) the video player (VP), (B) the video cube (VC), and (C) the Activity Cube (AC). The Activity Cube is the central element of Viz-A-Vis. We did not include the Activity Table for this study for two reasons. First, there is a natural progression between the video player, the video cube, and the Activity Cube. The Activity Table is an abstraction that naturally follows the Activity Cube, but it would have added confounding effects to the study. We wanted to determine the effects of visualizing motion in its contextual space. The study is a straightforward comparison between VC and AC since the only change between the two conditions is the abstraction from color pixels to motion pixels. Second, the Activity Table adds considerable complexity to learning how to interpret Viz-A-Vis. Since we had limited time for training and testing, we simplified Viz-A-Vis to be the Activity Cube plus indexing to original video.

Twenty-four participants received training prior to embarking on five observational tasks: (1) describe events; (2) find the beginning of long events; (3) find short sporadic events; (4) count short motions; and (5) track individuals. We measured the participants' task performance and asked about their condition preference per sub-task. The study excludes culturally or behaviorally focused observation. It omits

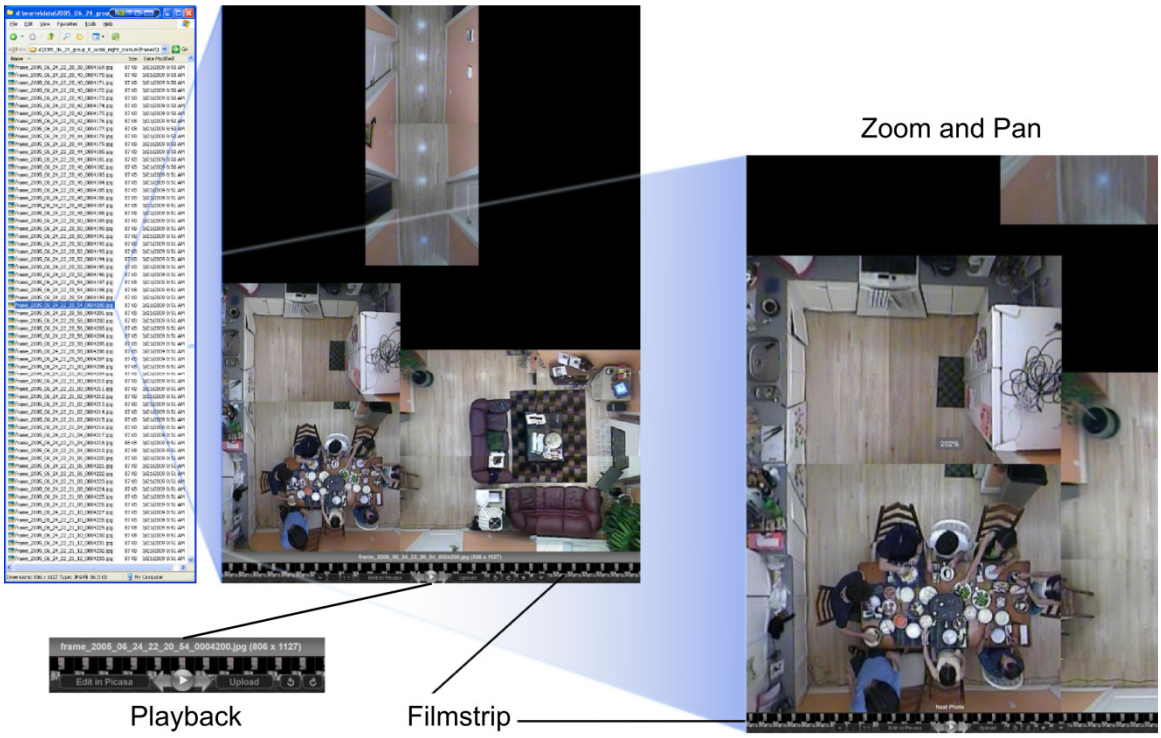


Figure 5.1: Performance user study condition A, the video player (VP) created from Windows Explorer and Google Picasa Image Viewer, a sample video frame from the sequence "having dinner," and the functional elements of VP: playback, filmstrip, zoom, and pan. Note that pixels map to physical locations.

inductive analysis such as classifying, synthesizing, or formalizing theories about activity.

5.2.1 *Conditions*

The three conditions are the (A) video player, (B) the video cube, and (C) the Activity Cube. The video player is the status quo. It is what is currently used most often for video playback and analysis. For our study, it is the control condition. The video cube is one of the state-of-the-art video visualization techniques. The Activity Cube is our innovation. It is the core of Viz-A-Vis.

5.2.1.1 Video Player

The first condition (A) is the control condition. The video player (VP) is a linear image sequence browser that supports traditional video playback. Figure 5.1 shows VP. It allows users to view the contents of individual frames sequentially at different speeds and arbitrarily by skipping forward or backward. Users can zoom-in and pan through individual frames in order to distinguish greater spatial detail. They can also navigate the video with a scrollbar that indicates the place in the sequence. Frames have individual time stamps and unique sequential frame numbers. Each frame is the result of unwarping, scaling, translating, rotating, stitching, and cropping ten overhead images so that pixel locations closely correspond to physical locations in the world. The resulting view resembles an orthographic projection of the space, but is really the stitching of multiple perspective projections. There exist multi-view geometry techniques for synthesizing an overhead view closer to a true panoptic orthography that are beyond the scope of this work. They are complex techniques with diminishing results.

Frames are full 24-bit color JPEG files with a recording rate of 2 frames per second (fps) and a resolution of 806 x 1127 pixels. This resolution is the maximum possible after stitching and cropping the ten original 320 x 240 pixel frames and leaving blank areas for the architectural spaces not covered. Figure 5.1 shows a sample video frame on top of VP and the labels of its functional elements.

Normally, video recording and playback requires at least 15 fps to appear continuous. Our animation of events does not look like continuous video playback. Played at 2 fps it looks like a slow time lapse. When played at speeds faster than 2 fps, it looks like typical time lapse. Users' typical interaction included playback speeds from

one frame per four or five seconds to several hundred frames per second. Speed depended on the task.

The video player navigation affords file search, sequential playback, filmstrip browsing, scrollbar timeline browsing, and end-point jumping. It renders individual full frames on a large window and a linear vicinity of thumbnails on the filmstrip. Users continuously control the speed of playback going forward or backward up to a factor of ten simply using the arrows on the keyboard. In other words, VP can play all the frames of an hour of video in 6 minutes. Through the filmstrip, users can time-lapse video skipping forward or backward up 20 frames at a time. The maximum time-lapse factor of VP in our computer is 120. It can play an hour of video in 30 seconds, but skips 19 frames out of 20 frames. Through the scrollbar, users can navigate directly to any point in the sequence. Furthermore, the scrollbar scrubs the thumbnail filmstrip of the frames. Finally, the video player can zoom and pan individual frames to view greater spatial detail.

We decided to keep the raw data format as individual images. Converting them to video produced prohibitively large files, lowered the quality of individual frames, and destabilized standard video players, such as Microsoft Media Player. Keeping the raw data as individual frames, we implemented the video player through a combination of Microsoft Explorer (Microsoft 2003) and Google Picasa Image Viewer (Google 2009). The data files are not hefty video files. Rather, they are lightweight image files. Thus, navigating through them with this combination of tools was the optimal solution without having to program it and without resorting to significantly more complex solutions, such as IrfanView (Skiljan 2009). Maintaining the data as separate image files maximized

portability, robustness, flexibility, and speed. All the other video player applications we tested were slower, more brittle, more blurred, and required significant extra work to transform images into video.

Strictly speaking, the purpose of Picasa Image Viewer is not video playback. The image-collection browser served well our purposes given that the video format is a sequence of JPEG images. Unfortunately, Picasa Image Viewer has two flaws for video playback. First, zoom and pan do not hold across frames when animating. Frames automatically return to the center of the screen and zoom to occupy the extents of the screen. The second flaw is a bug. A double click on the one-pixel edge of the filmstrip zooms the frame. The correct operation should be to direct the frame toward the target on the filmstrip. This was not a large issue during our experiments. When it occurred we explained it to participants and they moved on quickly. We had the option of using different tools or implementing a tool ourselves. We weighed stability, speed, and usability and voted for Picasa Image Viewer. We warned participants of these two flaws during their tutorial of VC.

5.2.1.2 The Video Cube

The second condition (B) is the video cube (VC) (Klein, Sloan et al. 2002). It maps time to space and presents the entire video sequence as a single three-dimensional structure. The cube maps image space to the horizontal dimensions of its base and time to its vertical dimension. Because images are essentially floor plans, the visual effect is that of a tall building, where each floor is a moment in time. A side cut of the "building" is a slice of space across time. A horizontal cut of the "building" is a video frame, a slice of

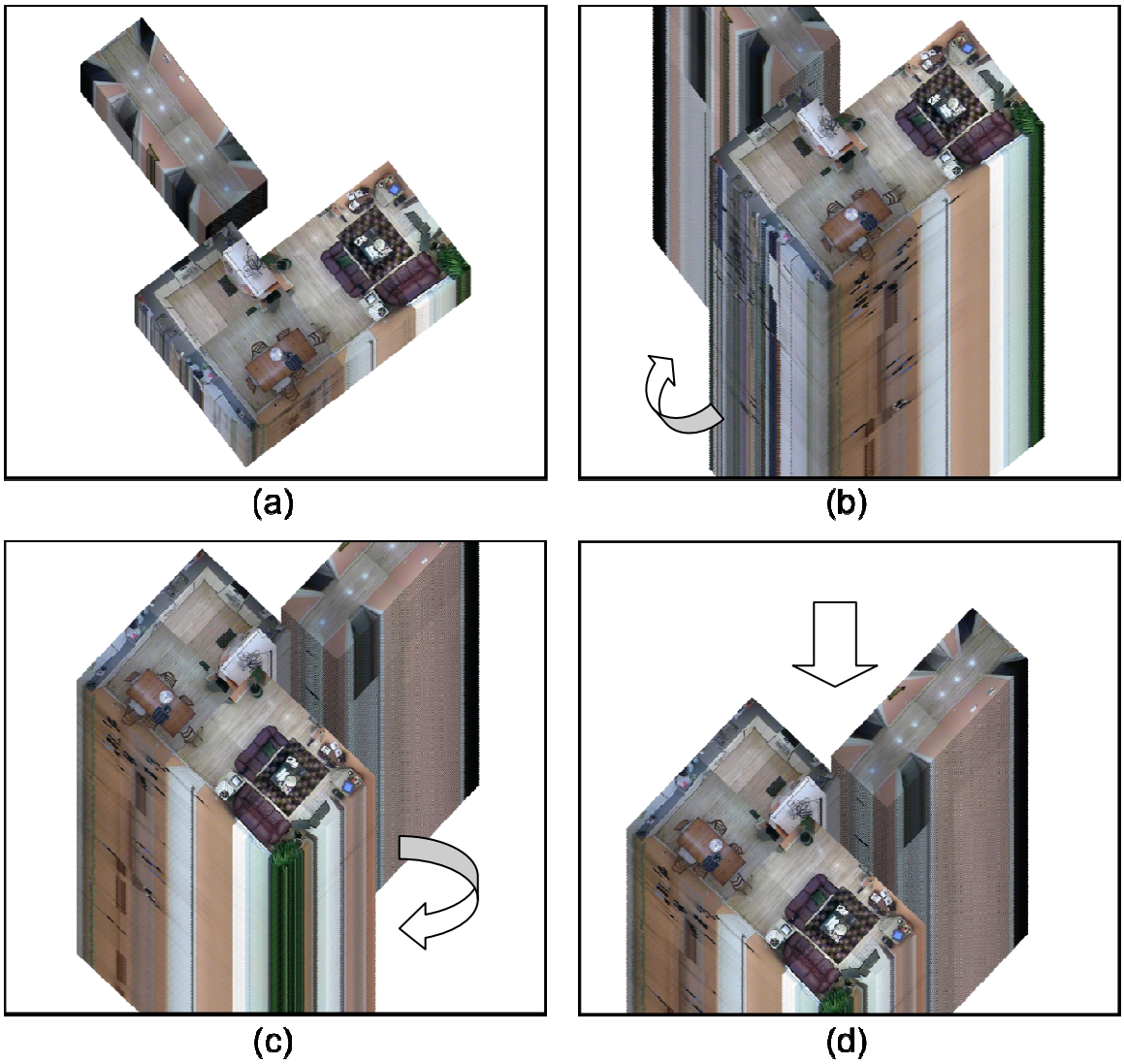


Figure 5.2: Performance user study condition B, the video cube (VC), created as a Ruby plugin for Google Sketchup and its navigation: (a) starting position; (b) orbiting up; (c) orbiting right; and (d) panning. Note changes on the sides of the cube indicating changes across time.

time across space, the traditional way of viewing video and experiencing reality. Figure 5.2 shows the video cube, orbiting, and panning.

The user can navigate and filter the cube to view different slices of space across time and different slices of time across space. As we will cover soon, unlike the Activity Cube, users can distinguish original pixels in VC, giving them the ability to determine the details of each image directly from VC. VC does not need a reification step.

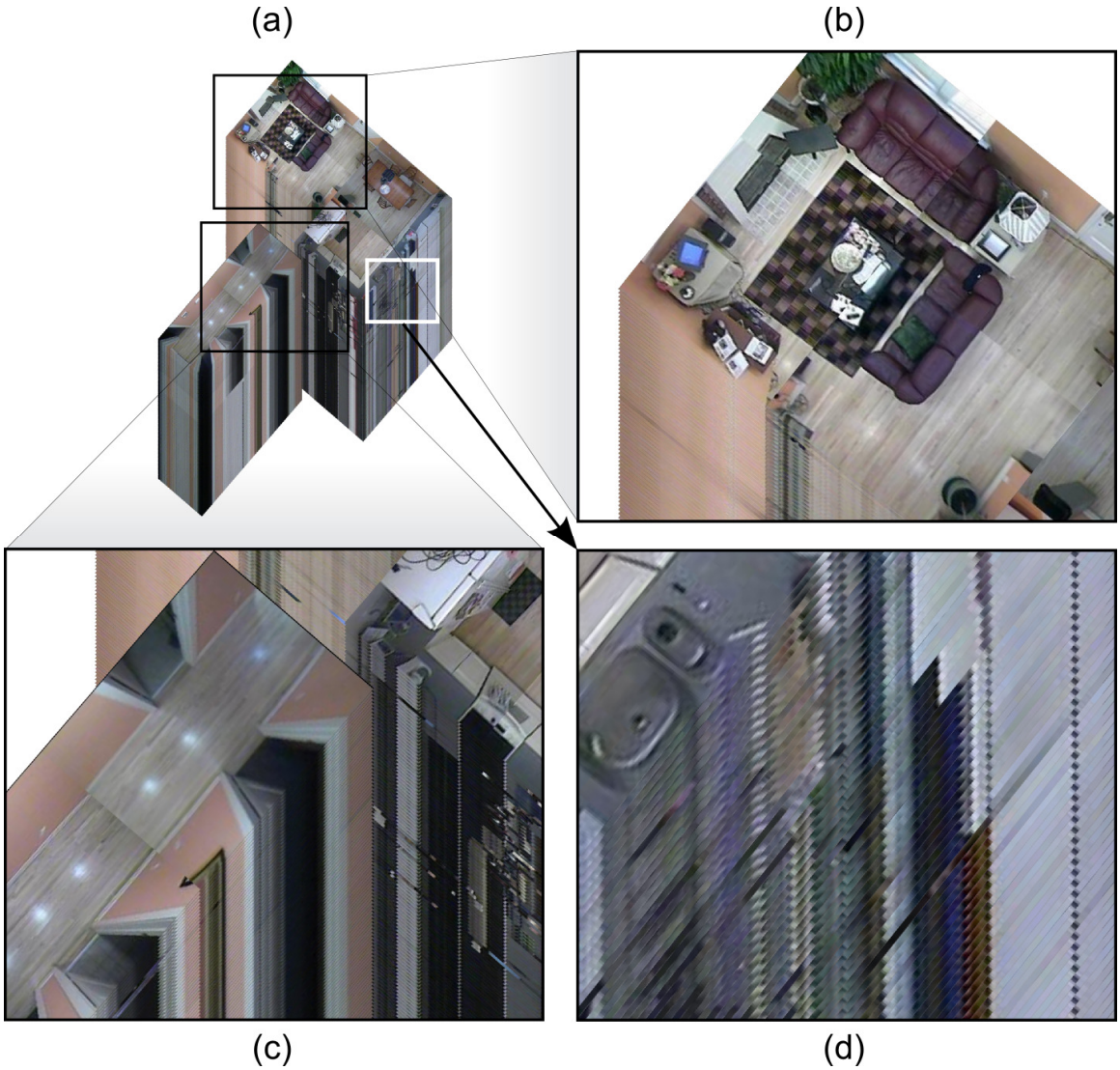


Figure 5.3: Three types of zooming: (a) original view; (b) directed zooming into living room; (c) centered zooming into center of original view, the hallway; and (d) windowed zooming to white window in original view, the sink.

The tools for navigating VC are orbit, pan, zoom, standard view, position camera, rotate camera, and change field of view. The tools for filtering VC are cutting and x-raying.

Orbiting circles the camera around the model while keeping it pointed at the center of the model. Ignoring the background, it has the visual effect of holding and rotating the model while looking at it.

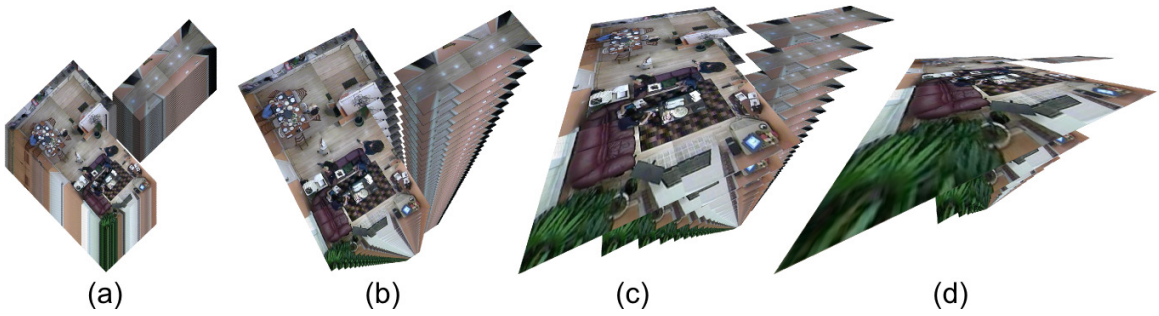


Figure 5.4: Video Cube Field of View (FOV): (a) parallel projection, FOV 0°; (b) FOV 45°; (c) FOV 90°; (d) FOV 120°.

Panning moves the camera on a plane perpendicular to the axis of projection in any direction. It keeps the orientation of the camera fixed, thus, the effect is that of holding and moving the model while looking forward.

Zooming moves the camera forward or backward along its axis of projection. The effect is similar to bringing the model closer or taking it away. There are four types of zooming: centered, directed, windowed, and extended. Centered zooming moves into and out from the center of the field of view (the screen). Directed zooming moves into and out from the mouse pointer. Windowed zooming defines a rectangular area that will be zoomed to occupy the full view. It is only a zoom-in operation. Finally, extended zooming moves the camera holding the direction of its axis of projection constant until the extents of the model occupy the entire field of view. It can be a zoom-in, if the model is far away, or a zoom-out, if the model is too close to the camera. Figure 5.3 shows the types of zooming.

Three-dimensional structures present three natural problems in two-dimensional projection and interaction: self-occlusion, ambiguity, and disorientation. To mitigate those problems, VC offers a number of navigation and filtering tools. Extended zooming is both a zooming and a reorienting tool. It allows users to recover their bearings,

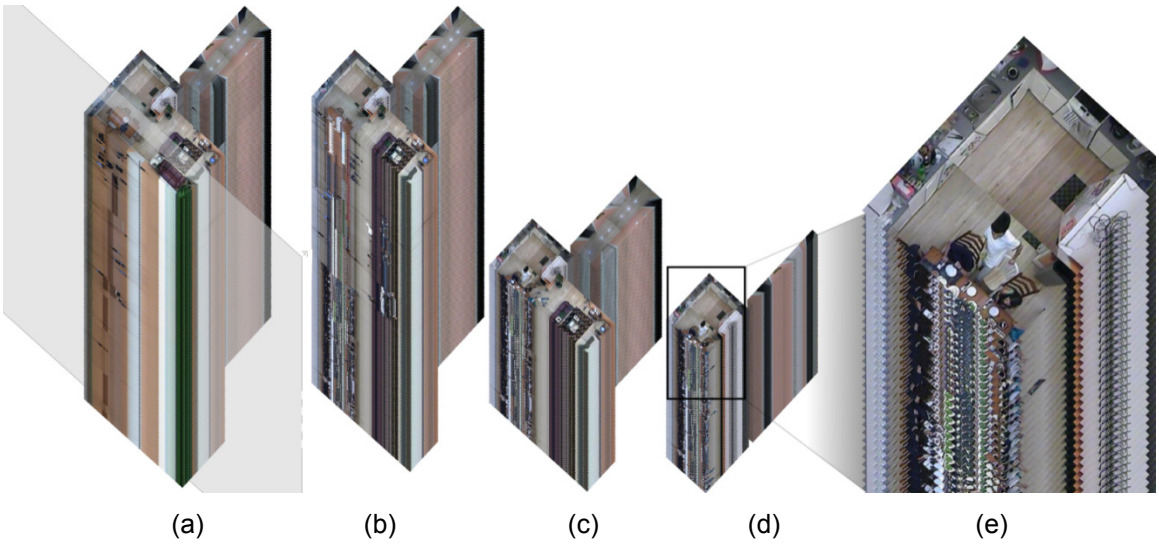


Figure 5.5: Cutting the video cube (VC): (a) section plane Y cut across dining room table (gray translucent plane); (b) Y Cut engaged; (c) Y-T cut engaged (note the T cut at the start of dinner); (d) X-Y-T cut; (e) zoom in on the Z-Y-T cut.

especially when they cannot see the model or are too close or too far away from the model to recognize their position with respect to it. The other navigational tools that also aim to mitigate disorientation are standard views, position camera, rotate camera, move camera, and change the field of view of camera. Position camera, rotate camera and move camera are also the first-person operations. They virtually place the user in the 3D view.

Standard views are top, front, back, left, right, and isometric projections. An isometric projection foreshortens the three coordinate axes equally so that the angle between any two is 120°. Figure 5.4 shows the standard views.

Position camera places the camera on the 3D coordinates pointed by a click. The effect is similar to flying towards a destination. Rotate camera rotates the axis of projection of the camera. The effect is similar to looking around from a first-person perspective. Move the camera repositions the center and orientation of the camera

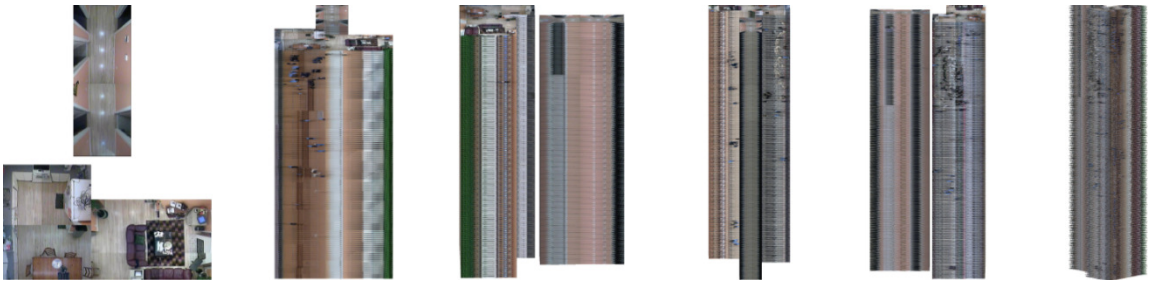


Figure 5.6: Video Cube (VC) standard views: (a) top; (b) front; (c) right; (d) back; (e) left; and (f) isometric.

according to a user-defined vector interactively and dynamically. The effect is that or walking around, similar to a first person shooter game.

The default field of view of the camera in VC is 0°. In other words, it is under parallel projection. With parallel projection, users have an equal view of the cube across its depth. Time and space do not get smaller as they get farther away from the camera. While this feature is useful under certain conditions and tasks, it is also somewhat disorienting under others. For example, it does not convey depth correctly. Users have the ability to change the angle of the field of view from 0° to 120°. In photographic imaging, the normal fields of view, those that feel natural, fall between 25° and 50°, with 46° being the most commonly cited (50 mm lens on a 35mm sensor) (Northey 1916). Figure 5.4 shows the effect of changing the field of view.

The second group of tools to help with the common problems of 3D interaction is the set of filtering tools. VC has three section planes that move parallel to each of the three dimensions and cut parallel to the other two dimensions (and perpendicular to the dimension it travels). The X-plane moves along the X-axis and creates a Y-T plane that cuts the cube at the position specified by X, perpendicular to x and parallel to Y-T. Figure 5.5 shows multiple cuts of the cube. Users can define the direction of the cut and hide the

(a) (b)

Figure 5.7: X-raying the video cube: (a) opaque and (b) translucent (x-rayed).

section plane while maintaining the cut active and moving it. In addition, users can activate multiple cuts simultaneously and define their own cuts, not necessarily at orthogonal angles. To simplify the study, we did not teach users to create their own cuts, to rotate the cuts freely, or to activate multiple cuts. Users could activate a single orthogonal cut.

Note that cutting along the T-axis is equivalent to watching the video player. The X and Y cuts were novel perspectives of space-time for all but three of the twenty-four participants of the study. The final tool for filtering is the x-ray tool. As its name implies, x-raying increases the translucency of individual layers, creating a ghost image of the inside of the cube. Figure 5.7 illustrates the effect of x-raying.

We implemented the video cube as a Ruby plug-in on top of Google Sketchup (Google 2009). Its performance, stability, functionality, and usability greatly surpassed our previous implementation in Matlab. Furthermore, Sketchup is a free distribution software platform where the authors of the Ruby plug-ins retain their intellectual property over the plug-in. Currently, the plug-in only loads the images and creates an index to the original files. In the future work section we will expand on the proposed functionality of

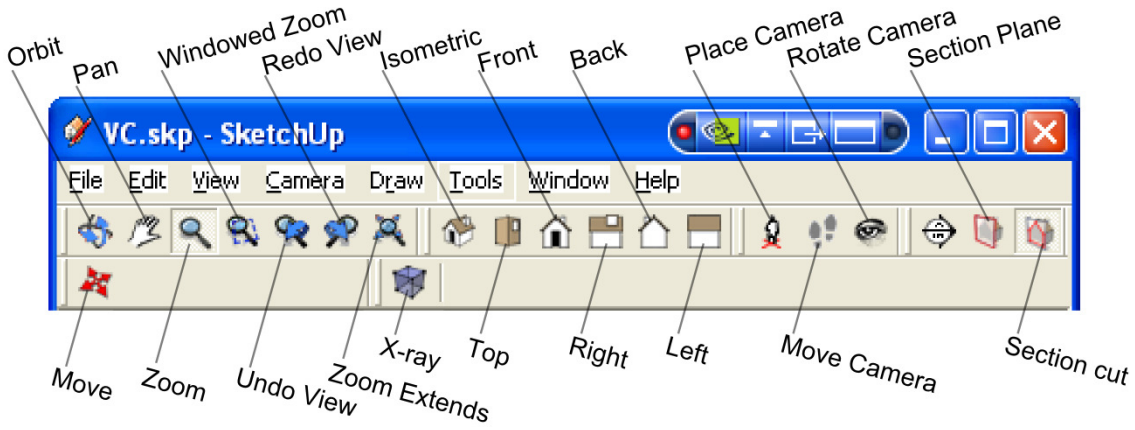


Figure 5.8: Google Sketchup interface: orbit, pan, zoom, windowed zoom, undo view, redo view, zoom extends, isometric view, top view, front view, right view, back view, left view, place camera, move camera, rotate camera, section plane, section cut, move, and x-ray.

VC as an integral part of Viz-A-Vis. Figure 5.8 shows the Google Sketchup interface and functionality. For computational efficiency, we sample the raw data at one frame per 20 seconds and provide indexing for frame-by-frame browsing through Picasa Image Viewer.

5.2.1.3 Activity Cube

The third and final condition (C) is the Activity Cube (AC). The Activity Cube is the central overviewing and indexing tool of Viz-A-Vis. The design goals of AC were to provide an overview of activity, a rapid indexing to original video frames, and a new perspective on activity. When we measure the performance of AC it includes the 3D navigation and filtering of the Activity Cube plus the indexing and 2D reification by sequentially browsing original images. When we ask participants to rank the conditions, we ask them to rank the functionality of AC alone, without indexing and sequential image browsing. Our goal was to tease the two factors apart to determine the true extents of the functionality of AC.

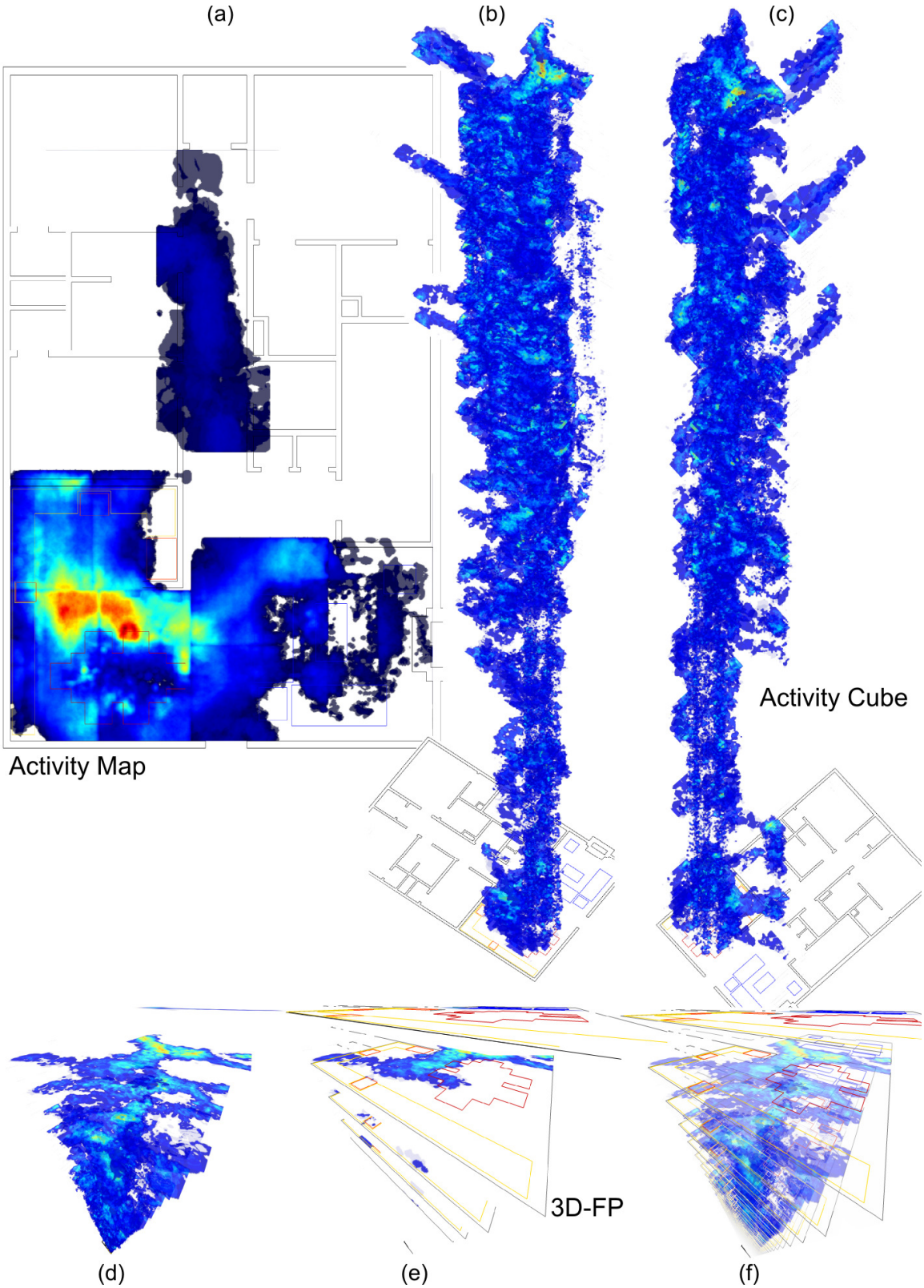


Figure 5.9: Performance user study third condition (C), Viz-A-Vis' the Activity Cube (AC) and its navigation: (a) the Activity Map (AM); (b) AC from the southwest; (c) AC from the southeast; (d) wide angle AC from the top without the 3D-floorplan (3D-FP); (e) AC with 3D-FP; (f) AC with 3D-FP and x-ray translucency.

The Activity Cube is similar to VC in that time maps to the vertical dimension of the cube. Section 4.2 gave a detailed explanation of the creation and manipulation of AC. Here we will describe one aspect of AC not mentioned in 4.2.

AC includes all the navigation and filtering functionality of VC, with one addition. AC is naturally very translucent and it lacks original pixels to give contextual information about the details of space or activity. While these features afford greater breath and length of overview and recognizing patterns of activity on a single view, they also produce greater ambiguity. There is ambiguity regarding where things are happening, in terms of spatial and temporal locations, and ambiguity in terms what is going on, who is generating the motion, and what objects are involved. We disambiguate spatial and temporal location with a 3D floor plan (3D-FP). The 3D-FP is an opaque repetition of the floor plan that segregates each layer and contextualizes the spatial location of activity. Users can activate and de-activate 3D-FP. The price of activating 3D-FP is the relative loss of depth. However, the rapid inclusion and exclusion of the 3D-FP or the x-raying of it provides an overview and context tool. Figure 5.9 shows a number of views resulting from navigating and filtering AC.

The computer vision component of AC runs on Matlab. We pre-compute the aggregate motion and store the results as Portable Network Graphic (PNG) files. We doubly map the aggregate motion to the black-to-blue-to-green-to-red heat-map and to the alpha channel (translucency) of the PNG file. The visualization is a Ruby plug-in for Sketchup. Again, the Ruby plug-in reads the PNG files and loads them to the 3D environment. It also creates an index table that the users can navigate back to the original frames.

5.2.2 *Data for Participant Analysis*

The data for participant analysis is a dinner party that started at 9:45 pm on June 24, 2005 (four years to the day of the defense of this thesis). Eight friends in their early 30s prepared food, had dinner, cleaned up, and played a board game. The gathering lasted four hours and fifteen minutes. Seven of the participants were Hispanic and one American. There were three females, three married couples, and two single males. The author and his wife were among the eight friends. They hosted the party at Georgia Tech's Aware Home. We split the recording of the dinner party into three scenes.

We show a different scene during each experimental condition to avoid learning effects on the data between conditions. The order in which we present the three scenes is always the same, regardless of the counterbalancing order. In other words, during session one a participant analyzes scene one, regardless of the condition, during session two, set two, and so on. The combinatorics of the distribution is as follows. Given 24 participants and 6 three-permutations of the 3 conditions, each permutation engaged four participants. These are the permutations: ABC, ACB, BCA, BAC, CAB, and CBA. Notice that each condition occupies a place eight times. For example, four people interacted with A first in ABC and four more in ACB. In short, each condition included each data set in each order eight times. Note that the total number of participant-condition pairings is 24 x 3 = 72.

The three scenes contain equivalent activities for the tasks. The tasks will receive a detailed description shortly. Briefly, the tasks are:

1. Describe the events
2. Find the beginning of:
    a. Dinner – 1st scene

b. Cleaning up – 2nd scene

c. Playing Cranium – 3rd scene

3. Search for bathroom visits

4. Count short motions:

a. Reaching the Raclette – 1st scene

b. Taking spoonfuls of ice cream – 2nd scene

c. Reaching for the Cranium™ board game – 3rd scene

5. Tracking individuals

The group of friends was conscious that the Aware Home camera system was active and they had all signed the appropriate consent forms. During the explanation of the data gathering, we clarified the purpose of the recording: to gather naturalistic data about human activity and use the data to visualize their activity, analyze it, and train computer vision systems. We asked participants to act as naturally as possible.

Once in the party, the eight participants immediately ignored the cameras altogether. We sketched two activities into the soirée: a raclette dinner and a game of Cranium™. A Swiss raclette is an electric grill that sits at the center of the table with raw ingredients around it. People place the ingredients on the grill and consume them when ready. Cranium is a board game where two or more teams compete to reach the end of the board. In order to advance, as a team, they must perform and deduce a number of tasks, some of which are very external, such as acting, singing, sculpting, and drawing, some are internal, such as spelling words backwards. Figure 5.10 shows the raclette and the game of Cranium™ we used on that day (Hasbro 2009).

Figure 5.10: Swiss raclette electric grill on the left and board game Cranium™ on the right.

The first scene lasts 62 minutes and contains 7307 frames. It includes the arrival of the group, the cutting of ingredients, the setting of the table, a number of conversations and interactions, the beginning of the raclette dinner, hundreds of instances of people reaching for the raclette, and two bathroom visits. The second scene lasts 41 minutes and contains 5191 frames. It includes the continuation of dinner, the end of dinner, the cleaning up of dinner, the setting up for dessert, having dessert, hundreds of instances of people taking a spoonful of ice cream, and two bathroom visits. The third partition lasts 94 minutes and contains 8382 frames. It includes the end of dessert, cleaning up, moving to the living room to play Cranium, setting up Cranium, explaining the rules, playing the game, hundreds of instances of people reaching for the board game, and one bathroom visit.

5.2.3 *Tasks*

In behavioral analysis, quantifying the *frequency*, *duration*, *latency*, and *percentage correct* are the primary descriptive tasks when focusing on behavior modification (Grant and Evans 1994). Since we are not focused on behavior

modification, we have translated these tasks to *describing, bounding, searching, counting*, and *tracking* based on Grant's description and on the low-level components of analytic activity in information visualization described in (Amar, Eagan et al. 2005). This list defines a practical number of tasks for the evaluation, not the possible tasks that users can perform when analyzing video. Before engaging in the five tasks, users received a theoretical and practical tutorial. We will describe the tutorial shortly.

5.2.3.1 Describing

Describing is the task of observing and verbalizing relevant features of activity captured in video. In the study, we started with a description to give participants a feel for interacting with the tool and with the data. At the end of the study we asked participants to rate the three tools based on how easy it is to describe events with each tool. We will describe the measures in detail shortly. During the study, we read a script aloud to be as clear and consistent as possible. The script is:

"Describe the video in at most 5 minutes. If possible:

1. Count the people visible
2. Identify people by their features
3. Determine where people go
4. Describe the rooms visible
5. Describe the furniture, appliances, and objects
6. Describe the spaces and places
7. Describe the activity
8. Describe interactions between individuals and objects, spaces, and other individuals

9. In general, make a story of the events that unfold

10. BALANCE LENGTH OF DESCRIPTION WITH DETAIL OF DESCRIPTION – Focus on what you find interesting.

Do you have any questions?"

5.2.3.2 Bounding

Bounding is the task of finding the frame of start and/or end of an activity. We asked participants to bound long activities. In contrast to searching, during bounding, users can find clues about the event in the surrounding video frames. For example, during the search of the start of dinner, users can track the plates on the table as they are being set up, the location of subjects around the house as they are gathering around the table, or people already sitting at the table and having dinner, if they skipped the beginning. Users can perform a type of logarithmic search around the event, where they use clues from the environment to determine if they have passed it or not and recursively hone in to it. The study script for bounding reads as follows:

> "Find the start of [dinner (scene 1); cleaning up (scene 2); the game (scene 3)] in at most 5 minutes. Before you begin, how do you define the start of [dinner; cleaning up; the game]? Do you have a strategy? What features will you look for? What tools will you use in your search? Do you have any questions?"

We ask participants to define, as concretely as possible, what they consider the start of the event to be. Our goal was to avoid confusion during the execution of the search. In our pilot runs, we determined that it is not always clear what "beginning of dinner" means. Some people define it as the moment when all are sitting at the table. Others define it as the first bite. It turns out that in the data, not all people are sitting at

the table when the first bite occurs. The same diverse conditions hold for all events in our three scenes and in human life. That is the nature of it. That is partly why automatic activity recognition is very complex. We had the option of defining the beginning of the activity, but we determined it detracted from the richness of the study. By letting users give the detailed definition, it allowed them to create hypothesis about the activities and test them according to the data.

We also asked participants to state their strategy, the features they will search in the video, and the tools they will use within each condition. Again, in our pilots we determined that people perform the search with greater clarity if they plan for it. Furthermore, people get lost in 3D interaction. During the study, we simulated expert behavior by providing contextual help to regain 3D bearings. In order to avoid providing strategic help, we needed to be clear about the strategy and tools participants would use.

There were 240 task instances (24 participants x 2 conditions in 3D x 5 tasks). There were approximately 54,720 seconds of task performance (24 participants x 2 conditions in 3D x 19 minutes of tasks x 60 seconds). Of the 240 tasks and 54,720 seconds of task performance, the evaluator provided help 28 times to 15 participants for a total of 203 seconds. We strived for the help to be verbal only and for it to direct towards a course of action stated by the participant prior to the start of the task. For example, if a participant were lost in navigation for a considerable time, we would recommend "why don't you try zoom extends?" We were very careful not to point to a direction the participant had not stated prior to embarking on the task. Of the 28 help instances, 15 were for AC and 13 for VC. The evaluator placed his hands on the controls three times, after repeated verbal and pointing attempts at the screen. We did not deduct the time used

for help from the time to task completion. We will discuss the contextual help further in section 6.4.

5.2.3.3 Searching

Searching is the task of locating in space and time instances of specific target actions, behaviors, or events. For this study, we chose target events that may leave subtle traces around it, but in general are unpredictable. We also chose events that have a definitive location and are relatively sparse in space and time. The fixed location served as an implicit clue for participants to use the affordances of the video cube and the Activity Cube, namely, vertical cutting at the location of interest. The particular search task we gave participants was to find bathroom visits. The script goes as follows:

> "Go to the start of the sequence and find as many of the bathroom visits in the sequence as you can in at most 5 minutes. A bathroom visit is an event where at least one person crosses the threshold of the bathroom door. The door does not need to close. When the bathroom is empty again, the visit is over. Do you have a strategy? What search tools will you use? How will you use them? Do you have any questions?"

5.2.3.4 Counting

Counting is enumerating the repetitions of a target action. We asked participants to count people reaching for the raclette, spoonfuls of ice cream, and people reaching for the game board in scenes 1, 2, and 3, respectively. The script is:

> "Go to start of [dinner (frame 4067); dessert (frame 10778); game (frame 14908)] and count people [Reaching out to Raclette; taking spoonfuls of ice cream; reaching for the game board] for at most 2 minutes.

"'A reach for the raclette occurs' when one or both hands of a person visually cross the threshold of the raclette. That is one reach. If the hand hovers over the raclette, it is still a single reach. Both hands must exit the threshold of the raclette before counting another reach. If multiple people reach for the raclette at the same time, you count each one individually. You only need to keep a single counter for everybody. In other words, you do not need to keep count for each individual. If the hand is obstructed, cut off from the field of view, pixilated, blurred, or unclear in any other way, make your best interpretation of the event. If the hand holds a utensil and the utensil clearly crosses the threshold of the raclette to manipulate food on it, count it as a reach.

"'Taking a spoonful of ice cream' is a person moving a hand with a spoon from the bowl to the mouth. You only need to keep a single counter for everybody. In other words, you do not need to keep count for each individual. If the hand or the spoon are obstructed, cut off from the field of view, pixilated, blurred, or unclear in any other way, make your best interpretation of the event.

"If one or both hands of a person visually cross the threshold of the game board, you count that as one 'game board reach'. If the hand hovers over the board, it is still a single reach. Both hands must exit the threshold of the board before counting another reach. If multiple people reach for the board at the same time, you count each one individually. You only need to keep a single counter for everybody. In other words, you do not need to keep count for each individual. If the hand is obstructed, cut off from the field of view, pixilated, blurred, or unclear

in any other way, make your best interpretation of the event. If the hand holds an object and the object clearly crosses the threshold of the board, count it as a reach. Do you have a strategy? What tools will you use? How will you use them? Do you have any questions?"

5.2.3.5 Tracking

Tracking is following the location and describing the actions of a target subject. It is a description refined to include only a single subject across space and time. This task is qualitatively different from description, not just quantitatively different. By asking participants to focus on a single user, the importance of identity increases significantly. This task highlights the lack of identity tracking in AC. The script for tracking reads:

"Go to frame [111, when all people arrive; 7724, the beginning of cleaning up after dinner; 13650, the end of dessert] choose an individual, identify the features of the individual, and track the individual (at most 2 minutes) – record level of description. If possible, describe:

1. The places and spaces the person visits
2. The objects the person interacts with
3. The actions the person performs
4. Appearance if it changes
5. The Interactions with other people
6. In general, tell the story of this person
7. BALANCE LENGTH OF DESCRIPTION WITH DETAIL OF DESCRIPTION – Focus on what you find interesting

Do you have a strategy? What tools will you use? How will you use them? Do you have any questions?"

5.2.3.6 Subtasks

Participants implicitly performed four subtasks during their execution of the five tasks explicitly stated above: (1) interacting; (2) short bounding; (3) overviewing; and (4) transitions targeting.

Interacting is performing the low-level control sequences that manipulate the data interface. We explicitly restrict the definition to simple clicks, drags, drops, and keyboard strikes. We exclude everything else, such as interpreting or understanding the data through the interface.

Short bounding is finding the extremes of short activities, for example, the start and end of visiting the bathroom, the fridge, or the sink, or the start and end of performing a charade during the game of cranium. Participants engaged in short bounding during the five explicit tasks. For example, they bounded the entrance and exit from the bathroom, a hand hovering on top of the raclette, or someone loading dishes to the dishwasher.

Overviewing is giving a shallow description of activity during the entire period and space of activity analysis. The description may be, for example, "people arrive, prepare food in the kitchen, eat in the dining room, and play a game in the living room." Participants performed overviewing during the primary tasks of describing, long bounding, searching, and tracking.

Transition targeting is finding the sequences where the subjects of observation move from occupying one space into occupying a different space. For example, the group

of eight friends moving into the dining room to eat dinner or moving into the living room to play a board game. Participants performed transition targeting when searching and bounding.

We measured the nine tasks through user self-report of condition preference for performing the tasks. In the next sections, we use the results from these measures to guide the overall result analysis and discussion.

5.2.4 *Performance Measures – Dependent Variables*

We utilize the following objective performance metrics: *time to task completion, precision, recall,* and *coverage*. *Time to task completion* (TTC) is the period between the start and end of a task, including its subtasks, and the user evaluation of results until the user is satisfied. *Precision* is the percentage of correct instances from the set of retrieved instances. *Recall* is the percentage of retrieved instances from the set of target instances in the original video. *Coverage* is the length of video traversed during the task. We fit these definitions for our research from the general information retrieval literature (Manning and Schütze 2002). Table 5.1 summarizes the objective performance metrics and the subjective user-evaluated performance metrics. Note that the order of the user-evaluated performance metrics comes from the order of the questionnaire at the end of the study.

We will cover the questionnaire shortly. Not all the measures apply to all the tasks. We measured precision, recall, coverage, and time to task completion for searching. We measure TTC and coverage for bounding. We measured precision, recall, and coverage for counting. We measured coverage for describing and tracking. All the tasks had an upper time bound to completion.

Table 5.1: Evaluation metrics of user tasks in the Viz-A-Vis user performance and preference study.

	Objective Performance				User-Evaluated Performance										
Task	Time to Task Completion	Precision	Recall	Coverage	1. Interacting	**2. Describing**	**3. Searching**	4. Short Bounding	5. Long Overviewing	6. Transition Targeting	**7. Counting**	**8. Long Bounding**	**9. Tracking**	10. Other	Interview
1. Describe				●	●	●		●	●					●	●
2. Bound (L)	●			●	●			●	●	●		●		●	●
3. Search	●	●	●	●	●		●	●	●	●				●	●
4. Count		●	●	●	●			●			●			●	●
5. Track				●	●			●	●				●	●	●

Using this language, our goals with Viz-A-Vis are: (1) to increase searching and bounding precision, recall, and coverage, thus lowering time to task completion; (2) to increase the view of activity across time in order to provide new overview vocabulary for the description and tracking of activity; (3) to provide a visual dictionary of behavior patterns across everyday episodes in order to facilitate new behavior pattern discovery; and (4) to improve the user experience by providing new perspectives of everyday life.

We define *behavior pattern discovery* as the task of systematically gathering and classifying evidence in the support of a theory connecting the causes (stimuli), effects (consequences), and observable features of newly witnessed behaviors.

The subjective measures we gather are user evaluation of the performance of each condition for the five tasks and four sub-tasks and for other tasks or sub-tasks we did not explicitly inquire. We delivered the questionnaire in writing and verbally at the end of the

study. We asked participants to justify their answers wherever they needed to do it. We also asked to distance the ranking to open greater opportunities and vocabulary for reflecting the answers. We asked how easy it is to: interact, describe, search, short bound, long overview, transition target, count, long bound, and track. The questionnaire is:

> "Please rank the three conditions (A – video player; B – video cube; C – Activity Cube) according to each category below and distance the ranking from 1, not a great difference, to 5, very different. Here is an example ranking how easy it is to (A) edit a word document, (B) send email, (C) create a website: B-1-A-5-C, meaning that, in my opinion, B is easier than A but by only a little bit and C is harder than A by a lot. Please, explain your answers whenever you feel the need. Do you have any questions?
>
> Conditions:
>
> A – Image sequence browser
>
> B – Video Cube
>
> C – Activity Cube
>
> Criteria for Ranking:
>
> 1. Easy to use (just the interaction – clicks, drags, buttons):
> 2. Easy to interpret activity (tell a story, describe events):
> 3. Easy to find short and sporadic events (bathroom visits):
> 4. Easy to determine the duration of a short event (how long in the bathroom):
> 5. Easy to get the global picture of activity over a long period of time (overview):
> 6. Easy to find a transition (group finished dinner and moves to living room):

7. Easy to count (for example, reaching for the grill or the game board):
8. Easy to find the start or end of a lasting activity (start of dinner):
9. Easy to track (people, objects, places):
10. Other categories you can think of, for ranking?"

We also gather the users' response to a hypothetical design task, where they choose the tools they will use and how they will use them. The script is:

"**Design an Application:**

"What would you use for video analysis? You can combine different tools for different parts of the analysis. Scenario: you are the designer of a large security system at an airport. You have installed overhead cameras over the entire floor plan of the airport and you want to design a system that will allow you to do forensic analysis of target events, that is, gather evidence of the causes and effects of the event. Of the three conditions, which tools would you use? Would you combine them? Would you add new features and capabilities?"

For the statistical analysis of the results, we summarized the data as mean ± standard error of the mean. We performed statistical analysis using Prism software Version 4.01. We conducted a one-way analysis of variance (ANOVA) for repeated measurements of the same variable. We used the Tukey multiple comparison test to conduct pair wise comparisons between means of each pair of groups. We considered differences at $P < 0.05$ to be statistically significant.

Finally, we gathered all the comments, critiques, and suggestions throughout the study. We group them into a number of categories using *focused coding* (Lofland and Lofland 1995). Focused coding is hypothesis driven. It concentrates on predefined

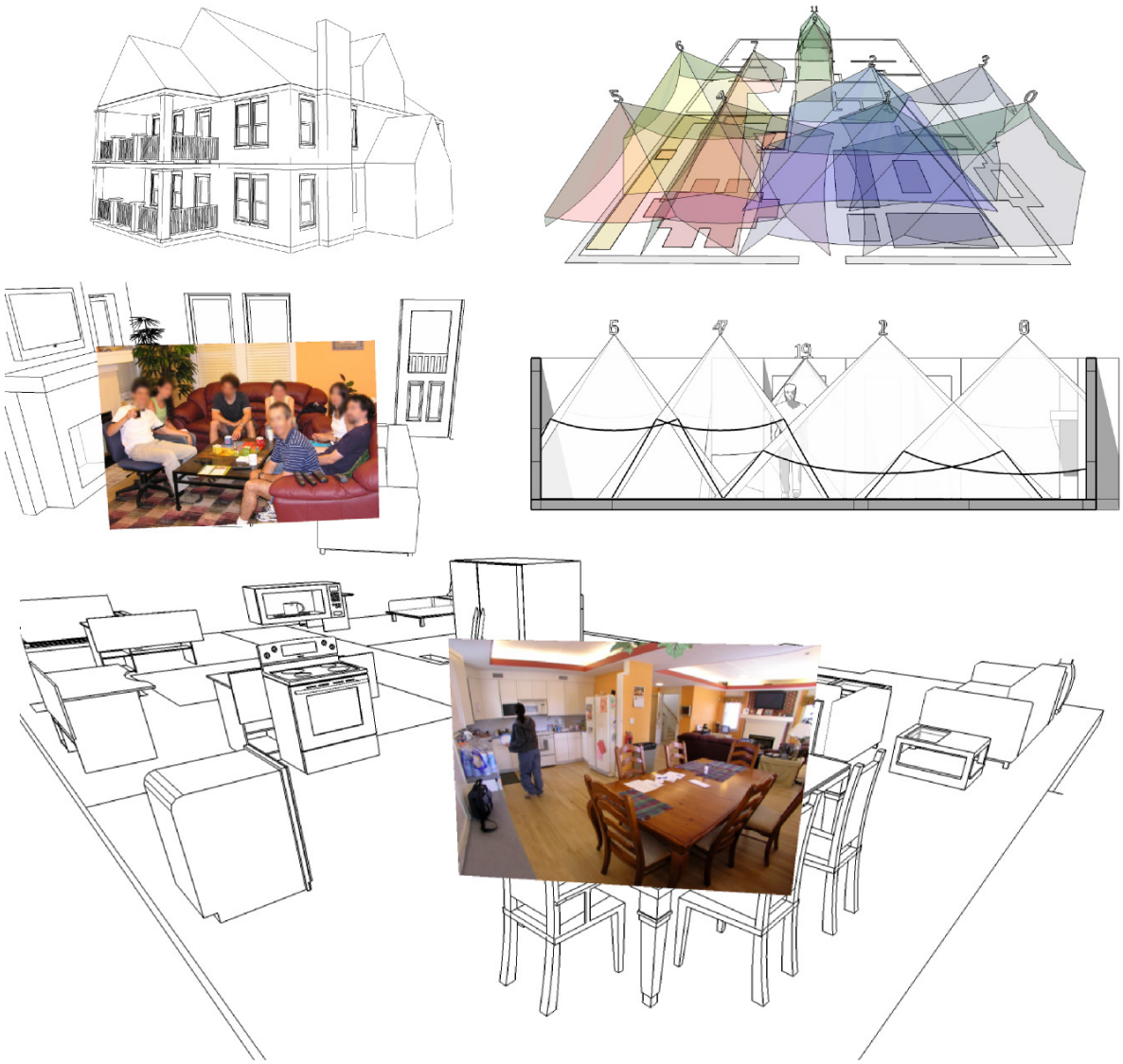


Figure 5.11: Viz-A-Vis snapshots from the tutorial for participants.

concepts relevant to a study's central research question. We grouped observations into concepts, concepts into concept-classes, and concept-classes into categories. We eliminated irrelevant categories and pruned scarcely substantiated concept-classes. Next, we merged categories until we had an irreducible structure. We explain the data based on this structure. We will present these results briefly.

5.2.5 *Participants and Testing Facility*

We recruited twenty-four participants, twenty from two undergraduate courses at Georgia Tech, Introduction to Human-Computer Interaction (Spring 2009 – Professor Gregory Abowd) and Introduction to Cognitive Science (Spring 2009 – Professor Rosa Arriaga). For participation in the study, both courses offered 1% extra credit for the semester grade as a means of compensating participants. We recruited four graduate students as well. Eighteen participants were male. The average age was 22.9 years. Most of the participants were computer science students. Most participants considered themselves experts at computer interaction and good programmers. Most cited at least some experience with data analysis and information visualization.

Most participants felt proficient 3D navigators, digital image, and video manipulators. On the other hand, most participants had never interacted with Picasa Image Viewer or Google Sketchup. About half had played Cranium™, but very few knew about projects at the Aware Home. We conducted a preliminary questionnaire to determine the level of training each participant would need. Figure 5.11 shows sample images from the training tutorial. Table 5.2 summarizes the statistics gathered from the questionnaire below.

"Initial Questionnaire

1. Initials:
2. Gender: M__ F__
3. Age:
4. Major:
5. Academic Year:
6. Experience with general computer interaction: 1 – 2 – 3 – 4 – 5

   (1. Novice. 5. Expert)
7. Experience with computer programming: 1 – 2 – 3 – 4 – 5
8. Experience with data analysis: 1 – 2 – 3 – 4 – 5
9. Experience with information visualization: 1 – 2 – 3 – 4 – 5
10. Experience with 3D navigation: 1 – 2 – 3 – 4 – 5
11. Experience with digital imaging: 1 – 2 – 3 – 4 – 5
12. Experience with digital video: 1 – 2 – 3 – 4 – 5
13. Experience with Picasa Image Viewer: 1 – 2 – 3 – 4 – 5
14. Experience with Google Sketchup: 1 – 2 – 3 – 4 – 5
15. Experience with board game Cranium: 1 – 2 – 3 – 4 – 5
16. Knowledge of Aware Home: 1 – 2 – 3 – 4 – 5

(1. Unknown, 5. Completed a project at the Aware Home)

Do you have any questions?"

Table 5.2: Participant demographics and skills for Viz-A-Vis performance and preference user study.

Gender:	Male: 18	Female: 6
	Average	**Std. Dev.**
1. Age	22.88	3.55
2. Academic Year	4.33	2.08
3. Computer Interaction	4.58	0.65
4. Computer Programming	3.75	1.11
5. Data Analysis	2.92	0.88
6. Information Visualization	2.96	0.81
7. 3D Navigation	3.29	1.16
8. Digital Imaging	2.92	1.38
9. Digital Video	2.79	1.35
10. Picasa Image Viewer	1.75	1.11
11. Google Sketchup	1.38	0.82
12. Cranium	2.58	1.50
13. Aware Home	2.21	1.10

Figure 5.12: Usability laboratory, GVU Center, TSRB 216-A. Note the position of the video camera, the microphone, the two monitors, the annotated keyboard, and the annotated help sheet on the wall.

Given three counterbalanced conditions, we could recruit participants only in multiples of six. Twenty-four participants was the largest number we could test given our resources. The study was originally designed to run for 72 hours, but early in the run, we decided to allow participants variable training times. In the end, the study ran for over 80 hours. We took copious notes of interactions, gestures, comments, suggestions, and critiques. We recorded measurements of performance live. Finally, we videotaped all the interaction and analyzed the parts of the video that required greater attention. We preferred videotaping over event logging because people's hand gestures provided important cues for understanding their interactions with Viz-A-Vis. In all, the analysis of the study required more than 160 hours.

We trained the participants by presenting the theory behind the three conditions at the appropriate times. We presented the details of recording overhead video at the Aware

Home and creating the overhead view of the entire environment. Figure 5.11 shows snapshots of the tutorial. Then, we gave a hands-on tutorial of each condition. VP required an average of 3 minutes of training. Both VC and AC required from 10 to 90 minutes of training, with an average of 25 minutes. People's skill varied widely, even after the end of the study. We kept track of their skill by the number of assists they required. A five means no assists. A four means one assist. A three means two assists, and so on. The participants' skill distribution at the end of the study was one 1, three 2s, four 3s, seven 4s, and nine 5s. That is, nine people performed all the tasks completely independently.

We conducted the study at Georgia Tech's GVU usability laboratory, TSRB 216A. It is a study-ready laboratory. Our main reason for using it was the quietness it provided for the participants. It allowed us to control distractions. The computer in the study has an Intel Core 2 CPU running at 2.4 GHz. It has 3.25 GB of RAM and a high-end graphics card for 2009, the NVIDIA GeForce GTX 280 with a graphics clock of 600 MHz, a processor clock of 1296 MHz, and 1GB of GDDR3 RAM. We used two 1280 x 1024 19-inch monitors. Finally, we used a keyboard annotated with Sketchup shortcuts. Figure 5.12 shows the physical experimental setup.

5.3 Analysis & Results

In this section, we focus on the analysis of the metrics that determine the clearest benefits for the Activity Cube and support our thesis statement. Before we focus on the positive results, we will summarize the results where AC will clearly fail or underperform when compared to VP and VC. Figure 5.13 summarizes the results of the user assessment of performance in a radar plot. For the tasks of describing, counting, and tracking, the

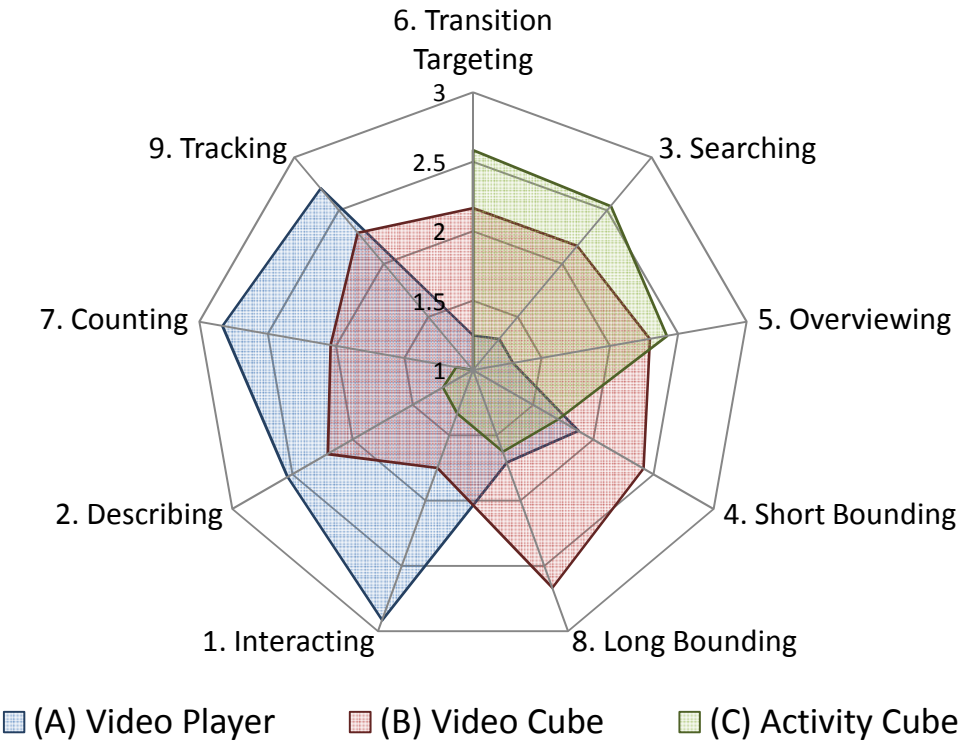


Figure 5.13: Average of the user assessment of the performance of each condition for the five tasks and four subtasks. The scale goes from 1 (worst) to 3 (best). The radar graph is sorted clockwise on the ranking for the Activity Cube across the nine tasks.

video player is still the clear winner. We observed participants and collected their measures of performance through the ranking in the final questionnaire. We also measured coverage. We determined that with AC, most people would overview the entire dataset in little over a minute and then spend the remaining four minutes using sequential image browsing to describe the video or track an individual. Moreover, for tracking the individual, users hardly spent any time looking at AC. Our goal with these tasks was to determine if users would leverage from features in AC to describe or track individuals. AC does not facilitate the tracking of an individual in an environment with multiple people because of its ambiguity. The aggregate motion masquerades identity. The pattern of use we discovered, however, is overviewing when describing.

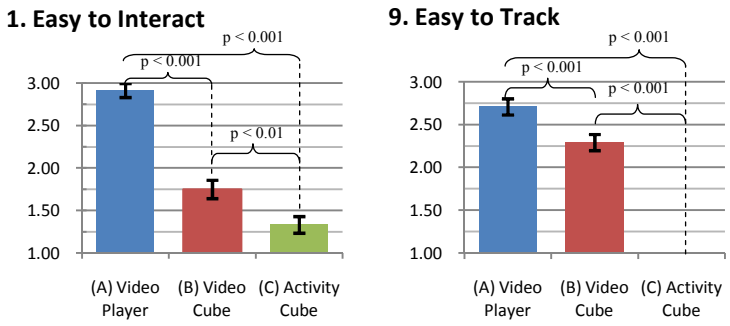


Figure 5.14: User assessment of condition performance for questions 1 and 9, where AC did not perform well.

In Figures 5.14 and 5.15, we show the average, standard error bars, and confidence value of user assessment for tracking, interacting, describing, and counting, questions 1, 2, 7, and 9 in the final questionnaire. We weigh user ranking: three points for first rank, two for second, and one for third. The ranking-plus-distance scheme we devised mainly prompted users to justify their ranking and we collected their rationale. These results, plus our observations during the study, are definitive. AC does not support tracking or counting and it is significantly harder to interact with AC than VP and somewhat harder than VC. VP is by far the easiest condition to interact with. The main reason cited for AC being harder to use than VC is the extra steps needed to manipulate AC in order to disambiguate it, namely using the 3D-FP.

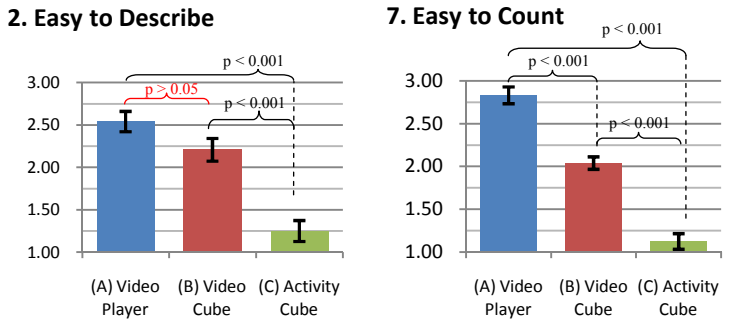


Figure 5.15: User assessment of each condition's support of counting and describing.

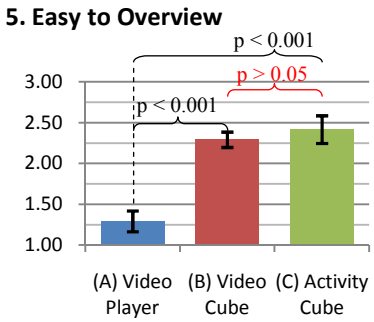


Figure 5.16: User assessment of each condition's capacity to overview video.

Question 5, overviewing, breaks up describing into detail and overview. AC does not support detailed description. For overview, on the other hand, it was ranked first, significantly over VP and slightly over VC. What are of interest are the reasons cited for the support of overviewing. VC supports overviewing by showing "everything in parallel." "What you see is not linear, it's parallel. [user places open hands in front of her, gestures to the space between them, and rotates an imaginary cube inside]." " I can scan [the sequence] with my eyes, rather than having to scroll." "I'm not worried about missing an event." "I get context with detail." The reasons cited for AC being better for overviewing are understanding patterns of movement, occupancy, and depth of view. "I get what's going on as a whole." "Flocking behavior is very clear. Everybody suddenly starts moving in the same direction." "It lets me see where the action is going on." "I can overview tendencies and trends. For example, I can see what areas are most popular." "[AC] allows you to view more data and zoom into what you are interested in." "I can quickly see the major activities."

Furthermore, when asked to design a hypothetical tool for forensic analysis in an airport, the only tool that users chose unanimously was AC for the task of overviewing. Some solutions included AC plus VC, others, AC plus VP.

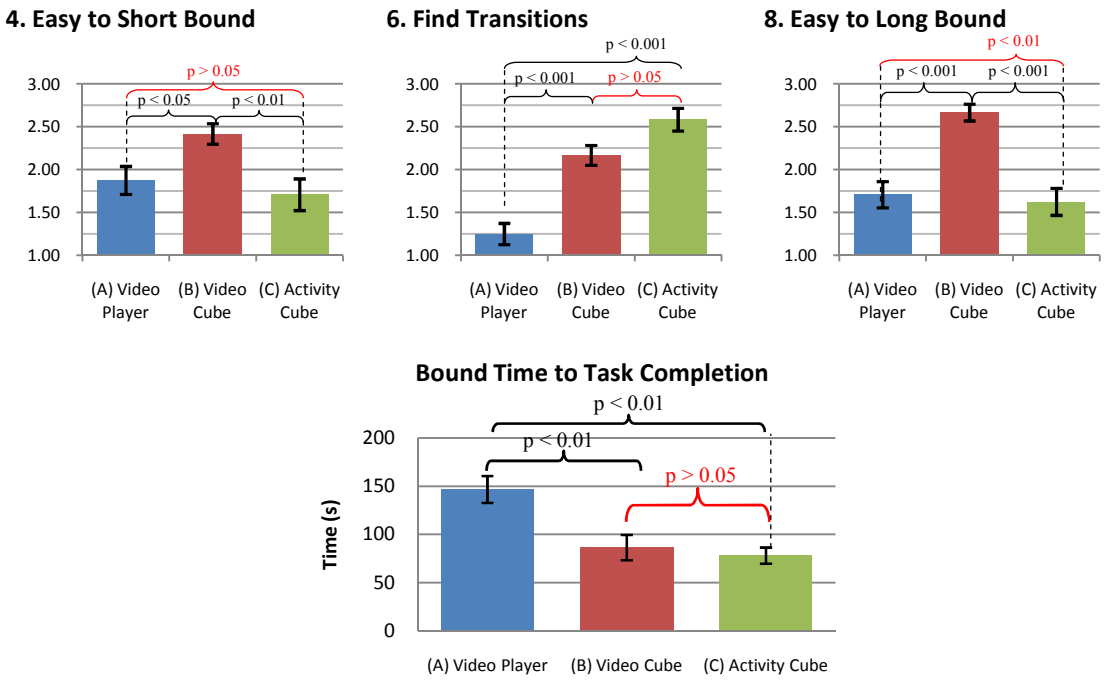


Figure 5.17: Top Row: User assessment of each condition's capacity to support bounding and sub-tasks of bounding. Bottom Row: User bounding performance, measured by their time-to-task completion in the bounding task.

Figure 5.17 shows the results of the user assessment of each condition's capacity to support the bounding task, finding temporal extremes of events, and the users' performance at the task, measured by time to task completion in seconds. The results are contradictory at first. Our hypothesis for this task was AC to perform better than VC and much better than VP. Our reasoning was that by carving the cube into the sub-spaces that contain motion, the user could focus on those targets more rapidly. Furthermore, large activities, like the start of dinner or the game, where everybody's behavior changes simultaneously, would have been easier to target with AC than VC and much easier than with VP. We discovered that it is not enough to see the general behavioral changes. When we asked users to bound activities, they were very precise in their answers. They returned with a frame number. AC points to the neighborhood of the boundary, but it does not take the user any further. To determine without ambiguity the beginning and end

of an activity, users need the detail of the activity. VC provides both the context of the entire time sequence, plus the detail of the pixels. It supported the task with greater elegance and users reported that. Nevertheless, time to task completion with AC was slightly faster. On the other hand, in the (possible) sub-task of finding spatial transitions of the entire group, AC stood out.

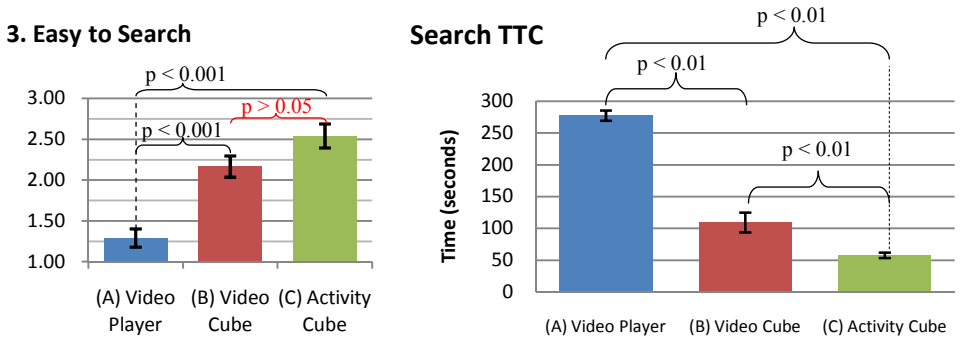


Figure 5.18: User assessment and performance of the three conditions (VP – video player, VC – video cube, and AC – Activity Cube) for the search task. Search precision and recall stood at 100% for all participants and all conditions. Search average coverage was 78% for VP and 100% for VC and AC.

Finally, for the task of searching, AC out-performed VP five-to-one and VC two-to-one in terms of time to task completion. Figure 5.18 shows search performance. In our study, all the participants found all the bathroom visits and all the events they pointed to were bathroom visits. In other words, everyone had perfect precision and recall. Furthermore, with VC and AC all participants covered the entire dataset. With VP the average coverage was only 78%. Had there been an instance of a target event in the remaining 22% of the dataset, recall would not have been perfect. The users' performance varied in time-to-task completion with an upper bound of 300 seconds. When using VP to search, we cut-off 16 out of 24 instances of search at 300 seconds. In other words, without the time limit, the difference would be greater. If we linearly interpolate, it would be an average eight-fold performance gain. Furthermore, the

sequence was not very long. Each dataset segment lasted about an hour. Five minutes to search for an occurrence in an hour of video is a reasonable task. When searching for occurrences over tens of hours of video, the tasks changes significantly. An interesting future experiment is testing the time to search completion of VP and AC as a function of video length. VP runs at O(n), proportional to the video length. Our hypothesis is that, as a function of video length, AC performs roughly in constant time, O(1) for the human operator!

In an extreme case, one of our participants performed the search task in two seconds. He rotated AC to its side, detected the strips running down the hallway, recognized the bathroom doorframe, pointed to it and said: "there is one and there is two." "Are you sure? Do you want to verify your answer?" "No, I'm sure." Indeed, in two seconds, he performed a 100% precision and recall search of two random instances of an event buried in over 7000 frames of video and in the process eliminated two possible false positives in the sequence!

We conclude this section with the results of the structured coding of the users comments and interviews. During their participation, we asked users to comment on the capturing infrastructure. Comments divided into coverage, occlusion, stitching artifacts, and detail. With respect to coverage, the two most cited problems were the blind spot on the foyer and the kitchen counters. Users accepted the lack of cameras in the bedrooms and bathrooms. It made sense to them that those spaces would be more private. The blind spot in the foyer annoyed participants. It generated a number of ambiguities regarding people's actions. It was not clear whether the subjects in the video were going to the media closet or outside the apartment. The kitchen counters, on the other hand, are spaces

were a great number of activities occur and the users of our study were constantly wishing they had coverage of the space on top of the counters. The participants suggested putting more cameras in areas where more dense activity occurred. Participants suggested having variable coverage over the space of observation. The other common and related suggestions participants shared was to have variable resolution over the space.

The artifacts of stitching were usually not a problem, until they directly obstructed the area of observation. For example, the game board was between four stitched areas, which meant the stitching cut-off many of the hand moves on top of it. Again, participants suggested dealing with the problem by placing more cameras centered on areas of interest and traffic.

Finally, participants mentioned the problems of self-occlusion, that is, people occluding their actions with parts of their body. Again, participants suggested placing cameras not just overhead, but in front of the areas of observation.

We consider these designs to be a function of the observation. Observational environments that require detailed and complete capture warrant the absolute and variable resolution coverage suggested by participants.

What we conclude from the comments of participants is that they keenly focused on observation and occlusions. Lack of detail, and being outside the field of view of the camera were conditions that truly frustrated participants. They really wanted to see and know everything. We conclude that is why they preferentially focused on the original images, rather than in abstracted images of AC. We did not give them the task of discovering patterns.

Participants touched on issues of privacy as well. Many mentioned that the blind spots, rather than being a lost opportunity for observation, were a haven for privacy, as long as the occupiers of the observed space knew where they were. Nevertheless, their comments made it clear that this observation infrastructure was not for everyone or for every environment. It is extremely intrusive and expensive to maintain and analyze.

At the end of the study, we asked two questions. Which of the three tools would participants use and how they would use them in the collection of forensic evidence at an airport? The second question was "In what other fields do you see potential for this technology?"

Participants suggested number of applications including tracking disabilities in the home, monitoring children with a sophisticated baby-cam and nanny-cam, observing social behavior, tracking of behavioral changes in classrooms with autistic children, and performing user studies, especially of ubiquitous and augmented reality applications, where body motion and physically observable behavior are part of the systems interaction and experience. Furthermore, participants suggested applications in leadership and group dynamic interventions for studying body language and gender roles.

In answering the design question, participants had a number of combinations and functional additions to the tool. What is paramount to this analysis is the fact that all participants included the Activity Cube as part of the tools of their application. Participants did not always choose the other two conditions, primarily because some participants considered that with easier interactive methods, the video cube completely subsumes the video player. Two participants went as far to describe the actual method: "the camera and the T-cut move in-step into the cube and the resulting animation will

look exactly like animating the video." Participants viewed AC differently. It was not redundant. It provided overview and pattern information that the other two could not. By "overview," participants meant that AC could look deep into temporal sequences, unobstructed by the self-occlusion problem of VC. By "pattern," participants referred to the ability of AC to visualize what is not easily detectable on either VC or VP. They used a number of words to describe it, for example, "clouds," "shapes," " blue patches," and "motion aggregates." "It shows where the action is and it shows the type of action by the amount of activity." "It gives some indication of activity. You lose who is doing what, but you get to see longer periods of time and where the activity happened." In short, participants regarded the motion aggregates as means of seeing deeper across time yet with a shallower lens. Furthermore, and of greater interest, participants saw AC as a new representation of activity that allowed them to see it differently and recognize patterns that would normally be obscured.

The most common combination for the airport scenario described was to combine the three tools for different tasks. Users would interact with AC when detecting activities in places or quantities out of the ordinary. The ideal use participants cited was tracking restricted zones, where unauthorized motion should not occur. In other words, AC would be used to track spaces where motion would be extremely sparse and limited to authorized times and personnel. We challenged users by proposing a simple motion detector system. Their response was to use the camera system not only to detect, but also to track and classify the unauthorized activity.

We close this section with our favorite quote.

[During training with VC] “Wow! I get the illusion that I see the scene from different perspectives. It feels like I’m moving the camera [that captured the data]! I feel I can see faces better when I look from the side. I know that’s not true, but I feel it anyway.”

5.4 Discussion

We will discuss seven important issues that we have not already covered in the description of the study or the analysis. First, it is hard to obtain objective measurements for this type of research, where participants have room to interpret the questions, for example, “what is the real beginning of dinner?” We segregated planning a search from executing the search to limit our assistance to execution related problems only and to avoid contradiction or confusion at search time. We could have included planning and defining in the time to task completion, but, in practice, it added roughly equal time to all three conditions. Users found the questions “do you have a strategy?” and “how do you define the beginning of dinner?” intriguing. Some users’ response was, “no, I’ll just zoom-through as fast as possible.” Others said, “yeah, I’ll do a logarithmic search looking for evidence of the event’s occurrence.” Yet others said, “not really… oh wait, maybe I can skip forward and look backwards, or maybe I can…” For this type of participant, our question influences their performance. We asked the exact same questions in the exact same order to all participants, but for some participants the questions influenced their behavior. Interestingly, it did not always improve their short-term performance, especially if they decided to explore different search strategies, rather than simply perform the search strategy they had had in mind to begin with.

Second, 3D navigation is an issue. Limiting the degrees of freedom of movement with sliders or buttons alleviates interaction and disorientation issues, but limits the possibilities of interaction. An ideal solution would be to have both types of interactive modes available to users, for example, panning and orbiting, both with the mouse and with hot-keys on the keyboard. The mouse provides greater freedom and the keyboard constrains disorientation. Sketchup does not support keyboard-based navigation. That is the main flaw of Sketchup for our tasks and one of the most cited possible improvements by users. We will implement keyboard navigation with Ruby on top of Sketchup for the next version of Viz-A-Vis. Regardless of the complexities of 3D interaction, we included all the tine penalties incurred during in 3D navigation when measuring time to task completion. In some instances, excluding rotations alone would have meant performance times of just a few seconds. Given the right perspective, the result is readily visible. We did measure the times independently to understand the impact of just 3D interaction, but the result is not very interesting. It is equivalent to eliminating the time it takes to navigate video from the total time to task completion of the search task. What is left but the recognition of the event in the target frame?

Third, the ideal for a study of this nature is to measure expert performance. Viz-A-Vis is a tool designed for expert analysts and 3D navigators. The target user population of Viz-A-Vis is people who will use the tool regularly and for possibly extended periods. In a future study, we would either recruit expert Sketchup users or train only people with intrinsic talent. We would not test this tool for participants who find 3D navigation very difficult. In our study, we had five such cases in this population. We compensated by extending their training time both at the beginning and during tasks. We repeated the

same lessons several times. During their performance of the tasks, they were the lowest ranking users. They received the greatest number of assists. This was a very frustrating, strenuous, and eye opening experience for the researcher. Mostly, it made him realize not all people experience our 4D reality equally.

Fourth, although Picasa Image Viewer has great features, it is not a state-of-the-art tool for video analysis. A number of research and commercial tools support text, audio, and video analysis, annotation, classification, and qualitative analysis in general, for example, (Mangold 2009; Max-Planck 2009; Muhr 2009; Noldus 2009). As we've stated before, our goal is to collect evidence from the data to conduct higher level analysis. All of the tools we researched focus on the higher-analytical part of the process, where the analyst compares and groups the collected evidence in order to build cases. Our specific goal is to support the task prior to that stage, where the observer collects the evidence. The commercial video analysis tools do not add to the tasks we studied. Picasa Image Viewer is representative of the observation, data, and evidence collection stage of the best video analysis tools.

Fifth, the Activity Cube shows a single, pre-computed level of aggregation. It does not support interactive temporal zoom-out into many hours, days, or weeks of temporal aggregates. We chose to show 20 second aggregates because it was the right granularity of the data for the tasks of this study. One of the greatest potential benefits of Viz-A-Vis may be the ability to zoom-out from a few seconds of aggregate video to a month of aggregate video. Due to time constraints, we chose to cut the temporal zooming feature of Viz-A-Vis. Of course, we can perform off-line temporal zooming with Matlab,

but interactive temporal zooming is the ultimate goal that we cut beyond the scope of this work.

Sixth, we purposely chose bathroom visits for this study. They are unpredictable, sporadic, isolated, and brief. They are unpredictable because they leave little trace before or after they occur. Therefore, evidence-based search is of little use. Participants tried counting people present, but that strategy did not work and they soon abandoned it. Large events leave obvious traces in the environment as they approach and pass. Participants can recursively hone into the event from the future and the past. Bathroom visits are sporadic because they do not occur often. They are isolated because they occupy a part of the space that is away from most of the activity. They are brief because they do not last more than a couple of minutes. Viz-A-Vis is a great bathroom-visit finder. We could do that with a simple switch on the door, or a motion detector in the bathroom. But what if the task it to find visits to the sink, the fridge, or the couch? The impractical answer with switches would be to place them anywhere. The answer with Viz-A-Vis is simply to take a look at that part of the space. Viz-A-Vis offers a flexible solution and, more interestingly, a solution open for exploration and discovery. Moreover, if the question is, "Does Viz-A-Vis support the search of any type of short event?" The answer is no. Repetitive short events, such as reaching for the raclette, are a blur for Viz-A-Vis. The key to the success of Viz-A-Vis as a search structure is the isolation of the target event, more than the other factors. Spatially isolated events stand out pre-attentively in aggregate motion space. Thus, Viz-A-Vis is a stupendous tool for finding isolated events, regardless of where or when they occur. In order to improve its ability as a search tool,

we need to artificially isolate target behaviors. Spatial zooming, as we have seen, is not enough. We will discuss our proposed methods in the future work section.

Finally, we discuss our choice of tasks. Remember, they are describing, bounding, searching, counting, and tracking. Given limited resources for the study, the goal was to obtain a representative sample of real practices from analytical communities. Our first approach was to experiment with the analysis ourselves, from an activity recognition and computational perception perspective. Literally, we used the visualization as a tool to build activity recognition and characterization algorithms for Tableau Machine. In activity recognition, the goal is to classify algorithmically raw sensor data into meaningful motions, actions, and histories of actions. At each level, the algorithm discovers statistical patterns of behavior and builds classifiers based on those statistics. The underlying assumption is that human life is a layered structure of micro-to-macro activities. For designers of perception algorithms, the first crucial and open-research step is to find relevant features in the data set. We find those features by observing and describing the data in human terms. Our first guide for the choice of tasks was the type of observation computational perception researchers engage in.

Next, we researched the environmental psychology and the behavioral analysis literature. In environmental psychology, researchers study the relationship between behavior and environment. For example in (Proshansky 1976), a technique called "environmental displays" methodically collects observational data under a number of categories including "free descriptions," "adjective checklists," "activity and mood checklists," "empathetic interpretations," and "social stereotypic cues." In behavioral analysis, the goal is to study behavior and the variables that influence behavior (Grant

and Evans 1994). Behavioral analysts strive to describe behavior with quantifiable precision in order to measure the effects of interventions. “Ate in five minutes” is a better description than “ate fast.” Behavior analysts measure behavior in terms of “frequency,” duration,” “latency,” and “percentage correct.” Frequency determines a count per unit of time. The count task accounts for frequency. Duration measures how long an episode lasts, from beginning to end. Duration maps directly to bounding. Describing, tracking, and searching came from a mixture of tasks and subtasks in computational perception, environmental psychology, and behavioral analysis. To the extent of our knowledge and experience, our choice of tasks is a representative sample of the typical tasks.

5.5 Conclusions & Contributions

In this chapter we evaluated Viz-A-Vis’ capacity to support observational analysis of behavior. With statistical significance, we objectively established its primary support task. Viz-A-Vis outperformed standard video playback five-to-one and the video cube two-to-one in searching brief, sporadic, unpredictable, and isolated events. In interviews with behavioral therapists, they report spending overwhelmingly most of their time searching for sparse target behaviors. Thus, we significantly improve performance for the task that matters most.

Second, both Viz-A-Vis and the video cube outperformed standard video playback two-to-one in bounding long events. Third, in the hypothetical design of a video forensics system, the only tool unanimously chosen for the system was the Activity Cube, the core unit of Viz-A-Vis. Users cited overviewing and discovering isolated patterns as the primary tasks for the Activity Cube.

To the best of our knowledge, these are the first valid claims regarding overhead video visualization for activity analysis. Other systems have presented their designs and short case studies, but the user study described in this chapter is the first rigorous evaluation of these technologies.

Finally, during this study a number of participants discovered several unexpected and potentially valuable activity patterns. Although discovery was not a task explicitly stated, participants stumbled upon a number of interesting activities. Although there is great value to such discoveries, it is hard to take standard measures of this task in a short study like the one presented in this chapter. Furthermore, the goals of the task were relatively generic. The short goals did not aggregate into a particular higher goal. In order to measure explicitly the impact on purposeful, innovative, and valuable discovery of activity patterns, we designed a domain expert evaluation. In the next chapter we present the design and the results of the study evaluating Viz-A-Vis' capacity to raise purposeful insight and discovery among domain experts.

CHAPTER 6

EVALUATING THE CAPACITY OF VIZ-A-VIS TO RAISE TASK-RELEVANT INSIGHT AND DISCOVERY AMONG DOMAIN-EXPERTS

In this chapter, we present the design, analysis, and results of a Viz-A-Vis intervention with domain experts. The goal of this study is to determine the capacity of Viz-A-Vis to raise interesting discoveries of activity patterns relevant to the design task of experienced architects. We observed two groups of architects during their design practice. Their task was to renovate the interior of the Aware Home given a number of constraints and requirements. Both sets of participants shared their comments, critiques, and suggestions in separate focus groups at the end of the study. Both groups received exposure to Viz-A-Vis, but only the second group before and during the design practice. We presented the tool to the first set of participants during their focus group.

Our main finding is that the Activity Table and the Activity Map engaged the architects in novel abstract conceptualizations of behavior in space and time that produced insightful and novel observations. Some of the keenest discoveries surprised not only the architects in this study, but also the generators of the behavioral data, the author and his wife. After some debate, analysis of evidence, and introspection, we recognized the veracity and value of the discoveries.

Section 6.1 recalls our research questions and states how this study relates to them. Section 6.2 presents the details of the design of the study and the rationale behind our decisions. Section 6.3 frames our choice of evaluation metrics and collection methods. Section 6.4 describes our analysis methodology and results. In section 6.5, there

is a discussion of the benefits and shortfalls of the analysis and results. The chapter closes with section 6.6, conclusions and contributions.

6.1 Research Questions

Remember the overall thesis of this work:

In the process of overhead video interpretation and analysis of activity, combining computer vision abstractions with information visualization techniques provides: (1) improved user task performance measured by time to task completion, precision, recall, coverage, and user assessment; (2) improved user experience measured by user preference; (3) increased user capacity to discover activity patterns; and (4) new opportunities for creative interpretation, experimentation, conversation, and reflection regarding everyday activities.

This study addresses the third thesis claim: increasing user capacity to discover activity patterns. The research question associated with this claim is:

Can vision-based data abstractions improve the information visualization interface as measured by analytical discovery of activity patterns? Can we provide a visual dictionary of behavior across everyday episodes in order to facilitate the systematic study of behavior and the discovery of behavioral patterns?

We define *behavior pattern discovery* as the task of systematically gathering and classifying evidence in the support of a theory connecting the causes (stimuli), effects (consequences), and observable features of newly witnessed behaviors.

6.2 Design of the Study

The purpose of the study is to measure the effects of introducing an activity visualization system in the information-gathering stage of architectural design, where

architects determine requirements, constraints, and behavioral patterns. These factors will guide their design. We conducted an intervention where we introduced Viz-A-Vis to a group of architects as part of their professional design practice. We determined the effects of the intervention as it relates to general changes to current practices. Specifically, we analyzed its potential to increase the architects' ability to discover activity patterns that will influence their design practice.

In this study, we investigate an architectural conversion. In their professional practice, architects are usually engaged in creations, not renovations. From this perspective, the study does not introduce the tool to the most common practice of Architecture. Nevertheless, given our resources, it was the most practical study. Regardless, we collected valuable data in the support of our conclusions. A deeper and more environmentally relevant study would include the introduction of the tool in the design project of new structures. In that case, Viz-A-Vis would support the discovery of patterns in existing similar spaces and the designer's tasks would include abstracting the patterns from its underlying context and reifying them to the new designs. In small scale, this is exactly the process we observed here.

We chose the participating architects because their practices include formulating design through the systematic study of the relationships between environment and behavior. Their current data gathering and analysis practices are labor intensive. For example, architects gather flow and occupancy by observing and counting or by interviewing and surveying. One relevant way they visualize the data is in geographic information systems, but the process is not streamlined and, more importantly, the variety and granularity of the data is very limited by the measuring lens: the observer. In this

Figure 6.1: Design session at the Aware Home with Group 1 without Viz-A-Vis.

domain, the two primary sub-goals of Viz-A-Vis are to streamline the capture and visualization process and to provide a capturing scope with greater coverage and higher granularity. In other words, we capture more types of data with greater resolution, automatic processing, and interactive visualizing. These improvements will expectedly afford a faster hypothesis generate-and-test cycle and a richer and broader space for hypothesis generation.

We observed two groups of architects during their design practice. The first group consisted of five doctoral architecture students and the second consisted of six. Their task was to renovate the interior public spaces of the Aware Home given a number of constraints and requirements as stipulated in writing and verbally by a fictional client. In both groups, each architect worked individually, but shared the space, the delivery of the requirements and the clients' answers to the questions posed by other architects from the same group.

Figure 6.2: Design session at the Aware Home with Group 2 with Viz-A-Vis.

The study had two sessions on separate days. The first sessions lasted between four and five hours and consisted of the design exercise. The second session lasted two hours and consisted of a focus group. The design and the focus group sessions took place, respectively, in the dining room and the living room of the second floor of the Aware Home. Figures 6.1 and 6.2 show the arrangement of participants during the design session of each group. Figure 6.3 shows the physical setup of the supporting infrastructure for the design: computers, printer, scanner, and large display screen.

The design sessions started with the delivery of the design program, the clients' presentation of their requirements, questions from the architects, sketching, second round of questions, refinements and delivery of presentations. For the second group the presentation of the requirements and current patterns included Viz-A-Vis. The focus groups started with a deeper presentation of Viz-A-Vis, a long round of questions and answers moderated by the author.

Figure 6.3: Technology support for design: computers, scanner, printer, and large display.

From the start, both groups were aware of the general goal of the study as we vaguely stated it: "to understand your current design practices and to determine the usability of a software tool aimed at supporting part of those practices." The first group was aware of the existence of the tool and they knew they would not meet it during their design session. They interacted with it during their focus group, where we introduced the theory and practice of activity capture and visualization in Viz-A-Vis. We showed them a number of episodes from daily living in the Aware Home and asked them to relate the visualizations back to their original design. We also motivated them to project how they could use the visual data in future designs. We observed, recorded, and transcribed the six hours of the design and focus group sessions. We collected their questions, comments, suggestions, and critiques, as well as the presentations of their design in visual, verbal, and textual media.

We started the second group design practice with a presentation and discussion of the workings and limitations of Viz-A-Vis. We visualized a number of episodes from the everyday life of the fictional client occupying the Aware Home during a period of nine

days and asked the participants to input queries into the system. “What does typical cooking look like?” “What does typical working in parallel look like?”

Participants delivered their queries verbally and we executed them on the visualization interface. We returned the results of the queries to all participants within a group and let them verbally guide the interactive views, allowing them to interpret the data. In order to limit the duration of the study, we delivered the queries through a dedicated technician instead of hands-on participant interaction. Hands-on interaction would have meant at least several hours of training and the proficiency would not have been that of the dedicated technician, who had hundreds of hours of experience. We were not testing the details of the interface in this study. Rather, we tested whether participants could interpret and utilize the results of the visualization to support their design task.

We digress here to put forward a potentially confounding factor we will discuss further in section 6.5. The author played four roles during the design session and one more role during the focus group. First, the author created the tool we tested. We did not hide this fact in order to motivate the participants by providing them with the real opportunity of having an impact on the tool by their participation. Second, the author and his wife played the clients. They also were the subjects of the collected data we visualized during the study. We modeled the fictional clients’ behavior closely based on the real life behavior of the author and his wife. Thus, the author was the subject of observation of the architects during the design exercise. Third, he was part of the team observing the architects during their practices. The observation included taking notes, photographs, and video recording. It did not include questions during the design. During the presentation of the designs of the architects, the author played both the role of the

client and the role of the observer when asking questions. Fourth, for the second group, the author played the role of the technician. He collected the queries, asked enough questions to eliminate any ambiguity, conducted the queries, and presented the results. During the focus group, the author moderated the discussion.

In designing this study, we had limited resources. Ideally, different people would play each role. In order to control this confounding factor, we took a number of steps. First, the study included five observers, three of whom are professional architects. Second, we carefully modeled and practiced playing the clients in order to deliver exactly the same descriptions and return equivalent answers to similar questions. Third, we carefully modeled the technician. He's task was only to deliver the results of the query. We carefully avoided including behavioral interpretations of the results. Fourth, we established an amicable environment from the start and we constantly encouraged criticism of our tools. Usually, people will give blunter criticism about a third party. Ideally, someone not related to the design of the tool evaluates it.

Returning to the description of the study, the second group had equal time limits to complete their design and the same deliverables. We presented to the group the results of the individual queries mid-way through their design and we collected their deliverables at the end. On a separate day, we conducted a focus group with emphasis on what worked, what did not work, what influenced their design, what was missing from the tool, and how they could use it to inform their future designs. Again, we observed, video recorded, transcribed, and analyzed the design and focus group sessions.

The primary purpose behind having two different groups was to understand and compare current practices with practices adopting the new technology. The first group

gave us a sample of current practices and, when presented with the new technology at the end of the study, a reflection on where and how they would incorporate the tool into their practices. The second group gave us a sample of how the practices integrate the tool without prior training. Furthermore, during their focus group, they reflected on what worked and what to incorporate into the tool in order to increase its integration into their practices and, ultimately, expand them. The secondary purpose behind the two-group intervention was to run a comparative analysis on the products of the design and find any difference between the conditions. The main obstacle to this analysis is that the internal confounding factors acquired over years of experience and innate talent will greatly outweigh any effect of our independent variable on the design product. To run a comparative study based on the product of the design would require a preliminary step of judging the participants' design proficiency and splitting the groups to be as balanced as possible. In our study, the two groups' design proficiency differed. The first group had a ten-year design experience on average. The second group had a five-year design experience on average. Participants naturally divided into the groups based on the availability on the days of each study.

We carefully defined the same task and schedule for both groups. They were in charge of renovating the Georgia Tech Aware Home's kitchen, dining room, living room, foyer, media closet, coat closet, south end of the main corridor, and balcony. Both groups had 30 minutes for initial data gathering, 120 minutes for initial sketches, 15 minutes for further data gathering, 60 minutes for final sketches and presentation material, and 5 minutes per architect for presentation. The total running time for the design sessions was 4 hours and 20 minutes, for the first session, and 4 hours and 50 minutes, for the second.

The extra time of the second session was due to the additional participant and the 25-minute presentation of Viz-A-Vis at the beginning of the session. For the data gathering sessions, we balanced the time of showing query results with Viz-A-Vis with the time of clients delivering their verbal accounts of their lifestyle. We kept it in the same time limits of 30 and 15 minutes each.

In the exercise, the Aware Home was a private house owned by a fictitious couple, *the clients*. The clients' requirements and limitations stipulated the design. The author and his wife played the clients. Both were present at both design sessions and delivered almost exactly the same description of their intentions and motivations. Furthermore, they stayed closely in character when answering the architects' questions, even when referring to detailed accounts of activity. The participants playing the clients were basing their portrayal on a detailed journaling of their real-life behaviors. Almost all descriptions of instances of activity and behavior were real occurrences only modified to fit the one aspect in which the client diverged from the players: wealth. The clients were wealthier.

Specifically, we modeled the clients' behaviors to match the behaviors of the author, his wife, and their guests during a nine-day data collection session between Friday, March 17 and Sunday, March 26, 2006. In that experiment, we collected a sample of everyday living data. At the time, the Hispanic couple was in their early 30s, both high-tech researchers. During their stay at the Aware Home, they cooked, had their meals together in the dining room and alone in the living room, watched television and movies in the living room, and worked in the living room and in the dining room during the weekdays. They filed their 2005 taxes on the morning Saturday, March 25, entertained a

Figure 6.4: Material for the design session at the Aware Home with both groups: floor plans, elevations, Sketchup and Auto-CAD models, and interior and exterior architectural photographs. Area of renovation marked under red box in floor plan at left.

different couple each Saturday evening, washed the dishes, swept the floor, vacuumed the carpets, went to the bathroom, slept in the master bedroom, and talked and browsed the web everywhere. In short, they performed everyday activities with a comfort equivalent to their own home. In addition, the author maintained the Aware Home computer vision system to ensure its stability. He summarized the events of each day in a journal and transcribed with detail a few samples of the nine-day period. Originally, we used the annotated data in the training of the learning sub-system of Tableau Machine, detailed in chapter 3.

For the architecture study, we created a fictional domestic interior design program and we distributed it to each participant on the day of the design session. Our participants were fully aware of the fictional nature of the design program.

Domestic interior design program

Prepared by client consultant: John Irritable

Client: Juan and Rosa Wizard

Distribution: Invited list of architects

February 21, 2009

Juan and Rosa Wizard have recently purchased an apartment on the second floor of a house on 10th and Center Streets, Atlanta. Based on your record of achievement as an architect, you are invited to submit a proposal with a title, no less than four power point slides and a text of no more than 300 words on the design of their new apartment. Your slides may include drawings, diagrams, sketches, or any other visual material that can convey succinctly and precisely your approach and proposal. We will consider any additional material of your choice, in any format or medium.

Preliminary description of the requirements of the client:

1. Juan and Rosa live intense and busy professional lives. Thus, they particularly value the time they spend together at home, even when they engage in different parallel activities.
2. Meals are very important to the culture of the family as a moment of togetherness. Thus, Juan and Rosa look towards structuring their meals as

special occasions and have used a dedicated dinning space for this purpose in the past.

3. Juan and Rosa enjoy entertaining friends at least twice a month. They offer their guests a meal and opportunities to talk, play games, watch movies, listen to music or jointly search YouTube, Picasa, and other sites of interest on the large living room screen. The mood dictates the duration of the social evening.
4. Both Juan and Rosa are avid computer users and digital media producers and consumers. Home theater, entertainment systems, computers, digital cameras, video equipment, screens and digital Art are always at hand in their home.
5. The family library is already extensive and growing. The collection of books is eclectic and covers a variety of subjects.
6. Juan and Rosa love music. In their daily lives, they wish for sound to be rich, clear, but not of high volume.

Please also find attached the following drawings in hard and soft (Sketchup and Auto CAD) copies:

1. Scaled plan of the apartment in its present form
2. Longitudinal scaled section
3. Transverse scaled section
4. Cut axonometric view of the interior
5. Southern external elevation of the building
6. Eastern external elevation of the building
7. Western external elevation of the building
8. Architectural photographs of the exterior and the interior of the home

9. Interior photographs of everyday living at the home, the clients and their friends

In Figure 6.4, we show samples of the floor plans, elevations, and interior design, exterior architecture, and lifestyle photographs we distributed to each architect.

In their description of how they use the space, the clients methodically mentioned or answered:

- 10th Street's noise and pollution prohibits the use of the balcony.
- The west cabinets in the kitchen are out of reach and are underutilized.
- There are too many cables, keyboards, computers, and electronic junk all over the place.
- Cooking is an enjoyable activity.
- Working occurs anytime and anywhere.
- The office is not the place where the working occurs.
- The couch in front of the fireplace and television is the place where we spend most of our relaxed time.

The groups received almost identical descriptions. When the participants asked similar questions, the clients' answers were similar as well.

6.3 Evaluation Metrics and Methodologies

Our primary focus of attention was the practice of design, not its product. The intrinsic confounding factors, acquired over years of exercised talent, greatly outweighed the single extrinsic independent variable of our short intervention. Furthermore, we did not have a practical way of balancing experience and talent between the groups. Even though we were not expecting results, Dr. John Peponis and the author conducted a

comparative analysis of the designs and found evidence of impact. We tabulated the "architectural movements" of each design. Architectural movements are the design's impact on the elements, features, and programs in the layout (Proshansky 1976; Whyte 1980; Van der Voordt and Wegen 2005). A program is the set of intended uses of a space together with the architectural affordances for that use. We will describe the analysis and its results in section 6.4.

Five researchers closely observed and analyzed each group's practice. The observers were John Peponis, Alice Vialard, Julie Brand, Natalia Landazuri, and Mario Romero. At the time of the study, the first observer was a professor of Architecture at Georgia Tech's College of Architecture and a member of this thesis' committee. The second and third observers were PhD students in the College of Architecture. The fourth observer was the author's wife, a postdoctoral fellow at Emory University's School of Medicine. The fifth observer, the author, was a PhD candidate at Georgia Tech's School of Interactive Computing.

We observed and took notes during the session. For the group that interacted with Viz-A-Vis, we paid special attention to their query-and-consumption process. The five observers were present at both design sessions. We also video recorded the entirety of both groups' design and focus group sessions. The author transcribed the video and synchronized it with the notes from the five observers. During the design exercise, we focused our observations on the participants' questions, comments, critiques, gestures, descriptions, and final presentations of their designs.

The three research architects and the author guided and observed the focus groups. We collected the participants' reflective evaluation based on any new information

provided by the tool, their interpretations and use, if any, of the visualizations, their critiques of the technology, and proposed future improvements and applications. We guided the discussion of the focus group to motivate the participants to think critically about the tool and its relation to their practice. Furthermore, we contained the discussion of the proposals to improve the tool to realistic upgrades within reasonable limits of computation and interaction.

Although Grounded Theory is arguably the most common method for qualitative analysis in interactive computing, for the analysis of our data, we used *focused coding* (Lofland and Lofland 1995). Grounded Theory is an inductive process of concept creation that bases its emergent theories on the open coding of the data (Glasser and Strauss 1967). Open coding presumes unbiased emergent classification of data. It is not hypothesis driven. Focused coding is hypothesis driven. It concentrates on predefined concepts relevant to a study's central research question.

We grouped observations into *concepts*, concepts into *concept-classes*, and concept-classes into *categories*. We eliminated irrelevant categories and pruned scarcely substantiated concept-classes. Next, we merged categories until we had an irreducible structure. Finally, we induced the resulting theory based on the relationships between the main categories.

To structure the presentation of our analysis graphically, we will use the following convention. When we introduce categories and concept-classes, we will use italics. We will always write categories in all capitals and concept-classes capitalized.

6.4 Analysis & Results

In this section, we present two analyses. The first is the comparative evaluation of the designs of each group based on the architectural movements of each design. The second is the qualitative analysis of the observation of the design experience. In order to present the analysis of the designs, we will briefly summarize each design and its proposed architectural movements.

Figures 6.5 to 6.16 summarize the design sketches and the architectural moves of each proposal. The summary of each design and its moves appear on the figures' notes. Table 6.1 summarizes all the designs and their movements. The label of each design is a two-digit number separated by a period. The first digit corresponds to the group and the second to the participant within the group in the order that they presented their proposal. For instance, Design 2.3 belongs to the third participant of the second group.

We classified architectural movements into the five categories that existed in this study: *integration of the balcony, segregation of the foyer, dedication of a media-centric space, visual linkage, and space bounding.* Integration of the balcony goes from the simple doorframe expansion and enclosing with windows of design 2.2, to the radical placement of the entire kitchen where the balcony is now, as in designs 2.3 and 2.6. Other designs that integrate the balcony are 1.3, 1.4, 1.5, and 2.4. Segregation of the foyer includes movements that extend the north wall of the living room or the east wall of the corridor or change the location of the entrance, as in designs 1.2, 1.3, 1.4, 1.5, 2.1, 2.3, 2.5, and 2.6. Dedication of a media-centric space is the creation of a space specifically dedicated to the consumption of media, as in designs 2.1, 2.3, 2.4, and 2.6.

Visual linkage refers to the internal and external opening of lines of sight through the simple inclusion of small windows to radical wall eradications. The most visually linking design is 1.4. The other visually linking designs are 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, 2.3, 2.4, 2.5, and 2.6.

Space Bounding is the movement that places architectural elements, such as walls and arcs, or furniture at the boundaries of a space, for example an island between the kitchen and the dining room. Space bounding and visual linkage are not mutually exclusive. The space bounding designs in our study are 1.1, 1.2, 1.3, 1.4, 1.5, 2.1, 2.3, 2.5 and 2.6.

In this analysis, the designs do not undergo judgment. The study lacks the tools and the experiment's design does not support conclusions regarding Viz-A-Vis from such analyses. We are not going to determine whether a particular movement is weak or strong, good or bad, appropriate or inadequate or whether a family of movements is correct. We are simply going to count and reflect on the differences as they relate to Viz-A-Vis.

From table 6.1 on page 186, we observe the following:

1. No participants in group 1 included media-centric space movements.

2. Four participants in group 2 included media-centric space movements.

This observation suggests an interesting phenomenon. Designers in group 2 became more sensitive to the need of space that is devoted to the consumption of media. Although participants in group 1 did address this need, they did no modify or reprogram a space to be mainly devoted to media. They simply presented variations based on the

existing layout. In both design sessions and in the design program we emphasized the clients' passion for media, yet only the second group acted strongly on it.

The difference may lie in Viz-A-Vis' presentation of behavioral patterns surrounding media consumption. In particular, the view of the typical Saturday morning showed the participants of the second group the myriad of activities the clients engage in while watching a movie. In Figure 6.18, the cells labeled "Saturdays" and "Taxes" contain data where the clients are watching a movie and doing multiple activities simultaneously. "Taxes" is a dramatic example where the clients filed their taxes online while watching Spiderman II. When they were ready to submit, the online tool imposed a charge the clients were not willing to pay and they re-did their taxes in paper-and-pencil at the west end of the dinner table, still watching movies. The west end of the table faces the television.

What is most important about this conclusion is that none of the architects in the first group generated this movement. It is not a definitive conclusion, but the extreme difference in the data supports it well.

In the next 11 pages we present a summary of each design, together with a brief explanation of the architectural moves and the justification for the moves in the words of each architect.

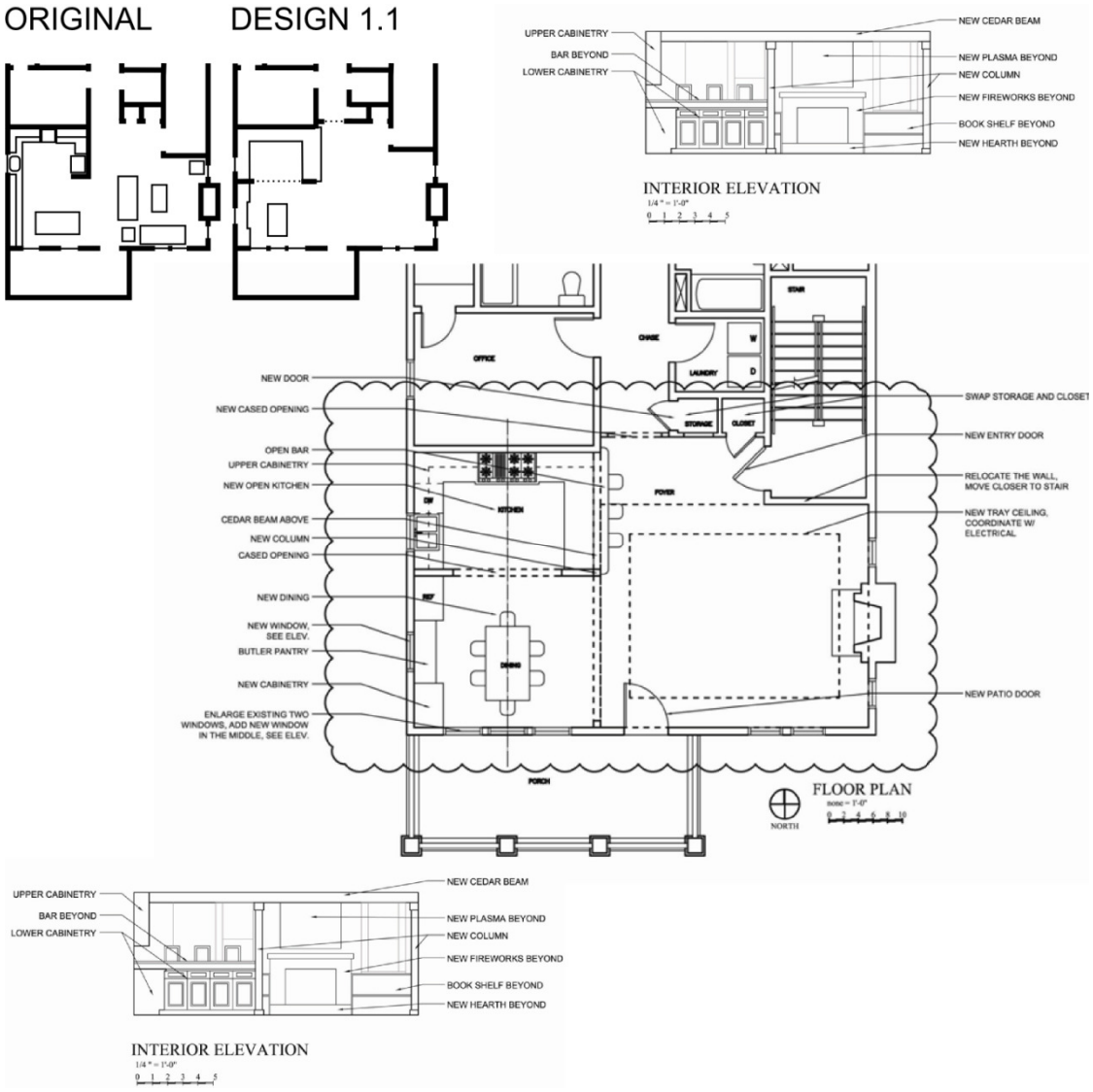


Figure 6.5: Design 1.1 title: "Wizard Residence Renovation." Architectural movements: visual linkage and space bounding. The architect removed the west kitchen wall between the kitchen and the foyer, included a counter and chairs to create a sitting surface. This move creates visual linkage. The architect placed a number of arcs, creating spatial bounding. Architect's statement: "Knock down a wall and create a great room."

Figure 6.6: Design 1.2 title: "stages for life." Architectural movements: segregation of the foyer, visual linkage, and space bounding. The architect replaced the west kitchen wall with a large counter top, extended the south wall of the foyer, moved and expanded the door to the balcony, and created a separation between the dining and the living room. Architect's statement: "Visibility supporting shared moments, even during different activities."

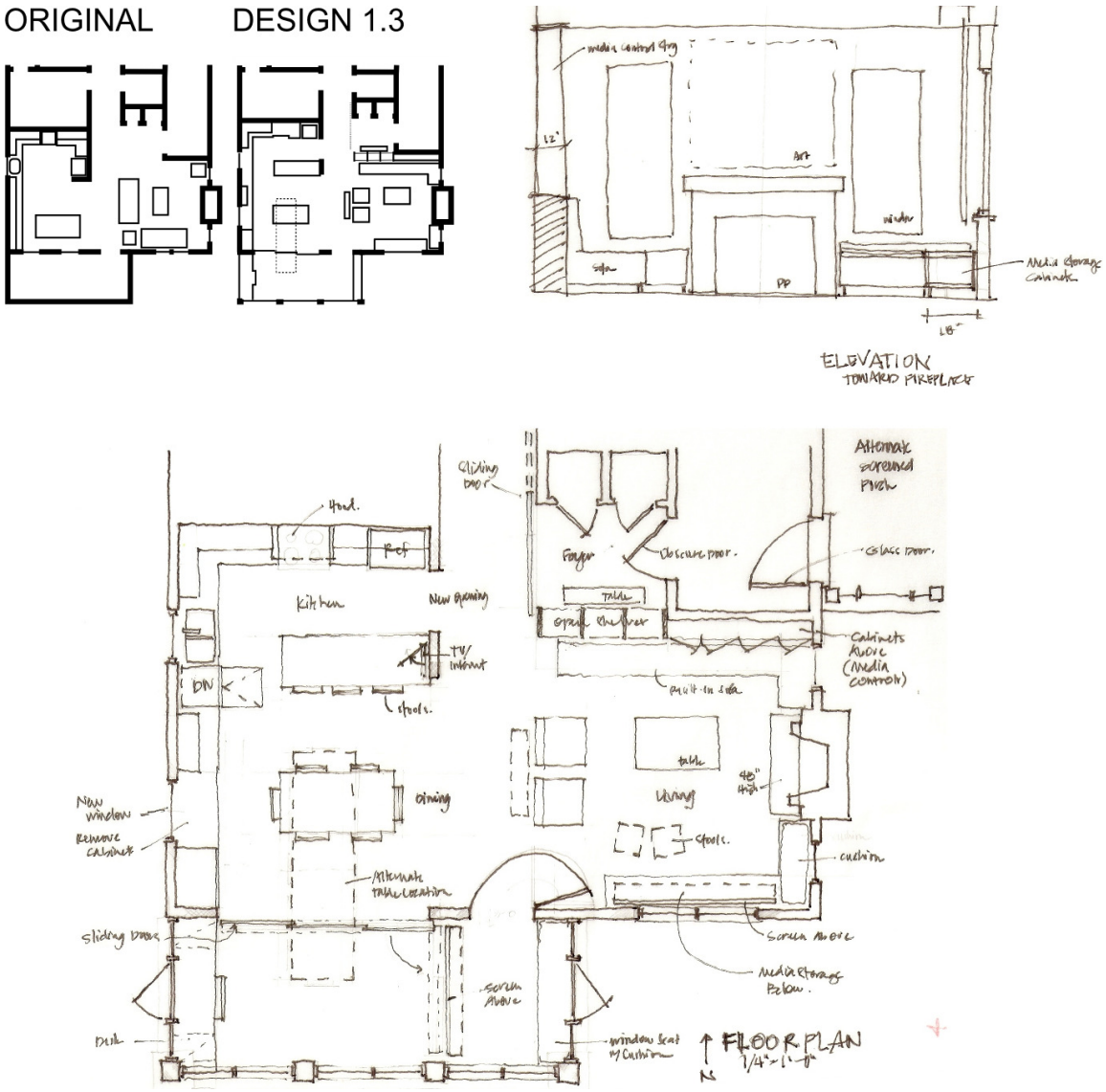


Figure 6.7: Design 1.3 title: "Sliding Doors." Architectural movements: integration of the balcony, segregation of the foyer, visual linkage, space bounding. The architect created a door on the west kitchen wall, moved the north kitchen wall a few inches further north, placed shelves on the south end of the foyer, and integrated the balcony into the interior space of the home. Architect's statement: "Expand social flexibility through visual and spatial integration."

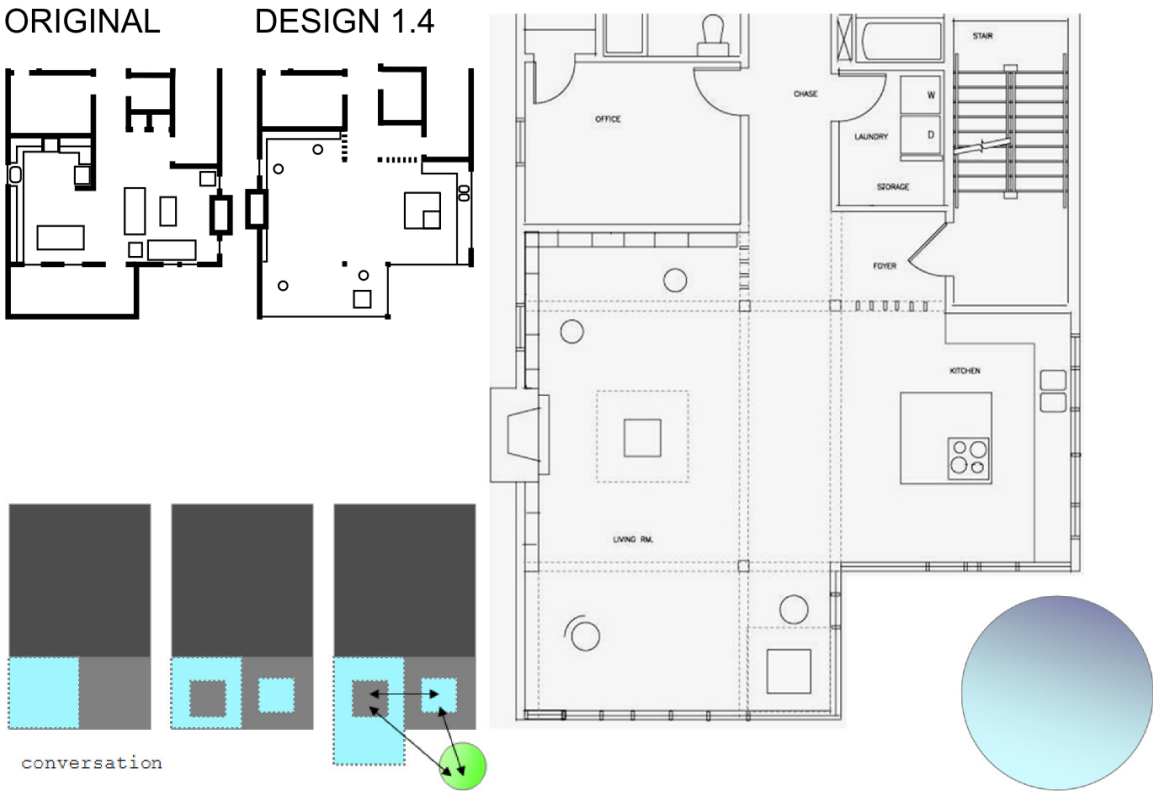


Figure 6.8: Design 1.4 title: "Space of conversation." Architectural movements: integration of the balcony, segregation of the foyer, visual linkage, and space bounding. The architect moved the kitchen to the southwest end of the house, moved the fireplace to the east wall, integrated the balcony into the interior space, visually integrated the outside tree, and placed bookshelves to create a bounded foyer. Architect's statement: "voice-sound-music-view: a democratic visibility."

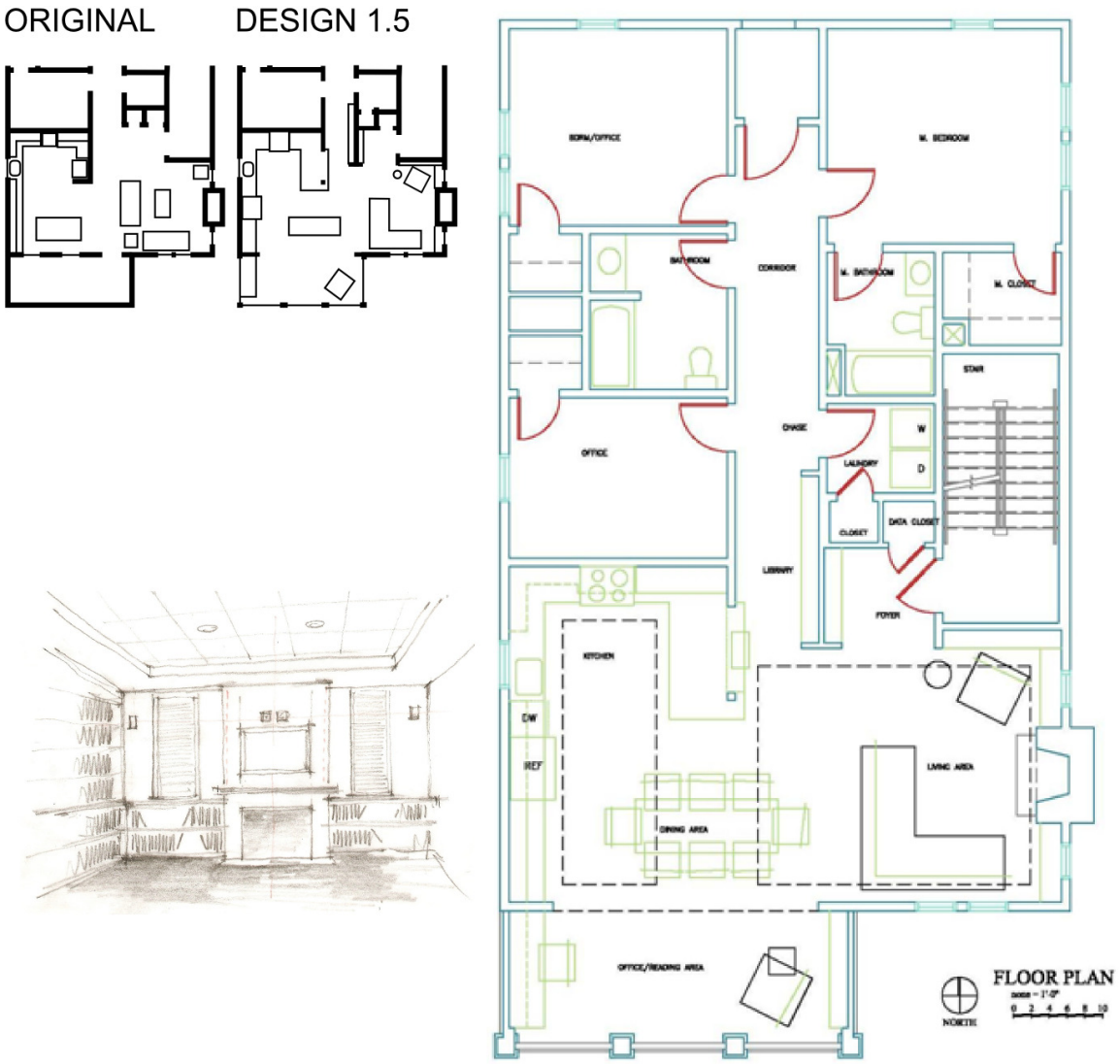


Figure 6.9: Design 1.5 title: "modern living as logic." Architectural movements: integration of the balcony, segregation of the foyer, visual linkage, space bounding. The architect replaced the west kitchen wall with a large counter, extended the east wall of the foyer, introduced the balcony into the interior space, and placed a number of shelves throughout. Architect's statement: "Simplicity and organization through the library and electronics. Dining as the central element for organization and focus."

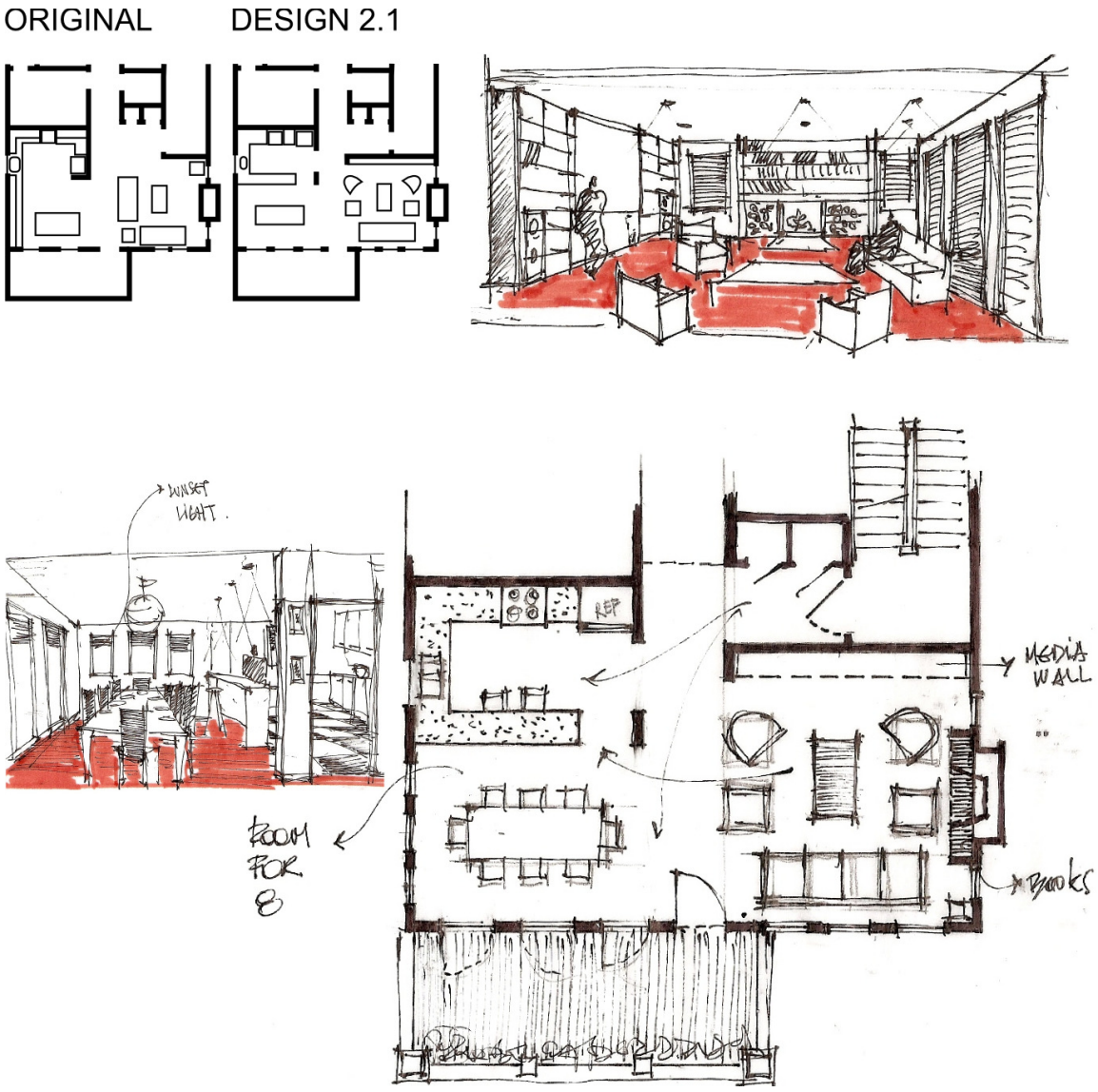


Figure 6.10: Design 2.1 title: "Single space – multiple space." Architectural movements: segregation of the foyer, dedication of a media-centric space, visual linkage, space bounding. Architect replaced the west kitchen wall with a large doorframe, added a counter between the kitchen and the dining room, added tall windows throughout the south wall, extended the south wall of the foyer, created a large projection media wall on the north wall of the living room, and created a living room oriented and dedicated for media consumption. Architect's statement: "Close kitchen by counter surface making dining space formally different. Open kitchen by removing wall. Switch focus between fireplace and wall, for books and media and projection, respectively."

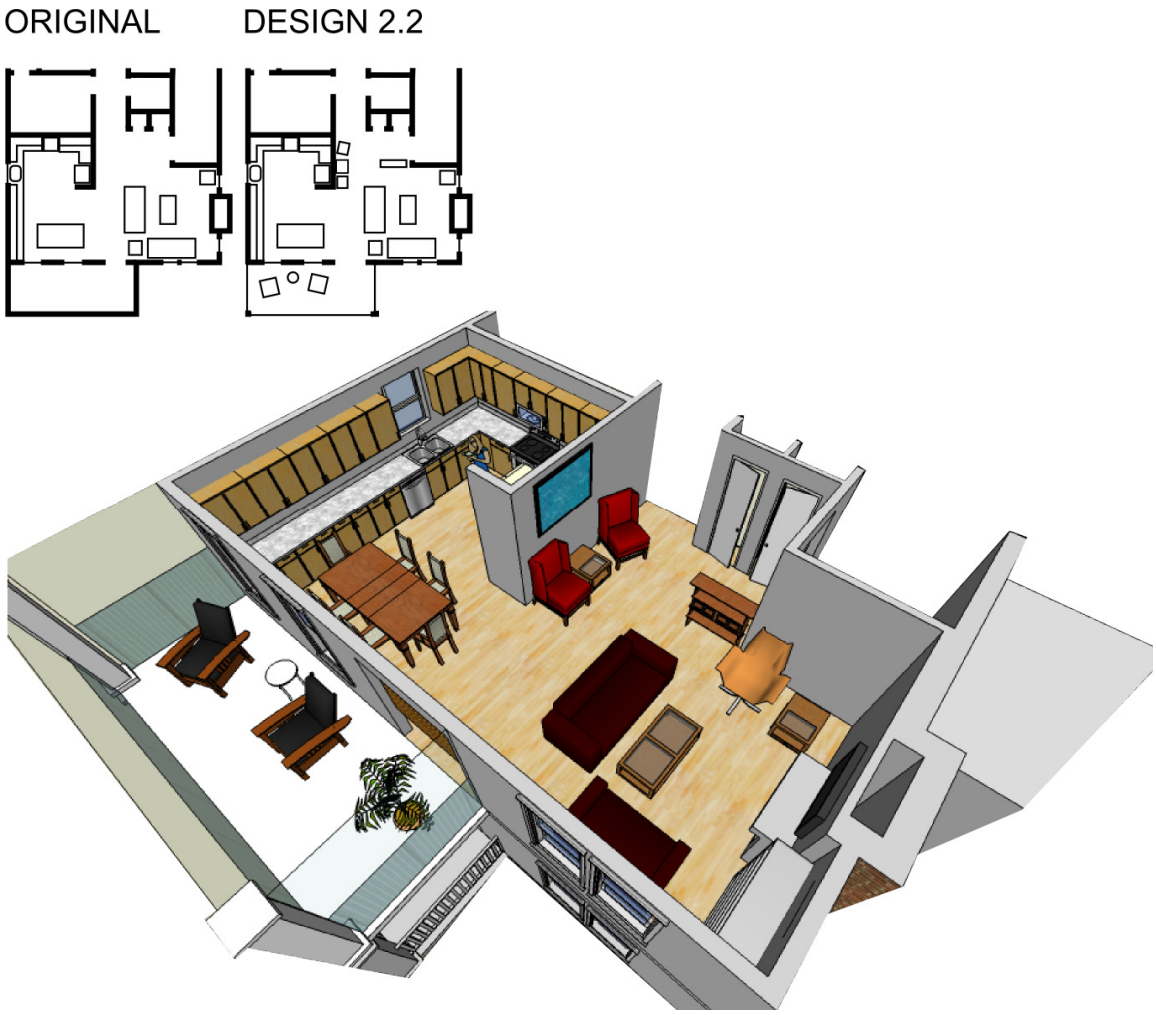


Figure 6.11: Design 2.2 title: "Living with Nature." Architectural movement: integration of the balcony. Architect expanded the door to the balcony and closed off the balcony with windows. Architect's statement: "Nature (lights and shadows). Foci of attention. Shared and separated spaces."

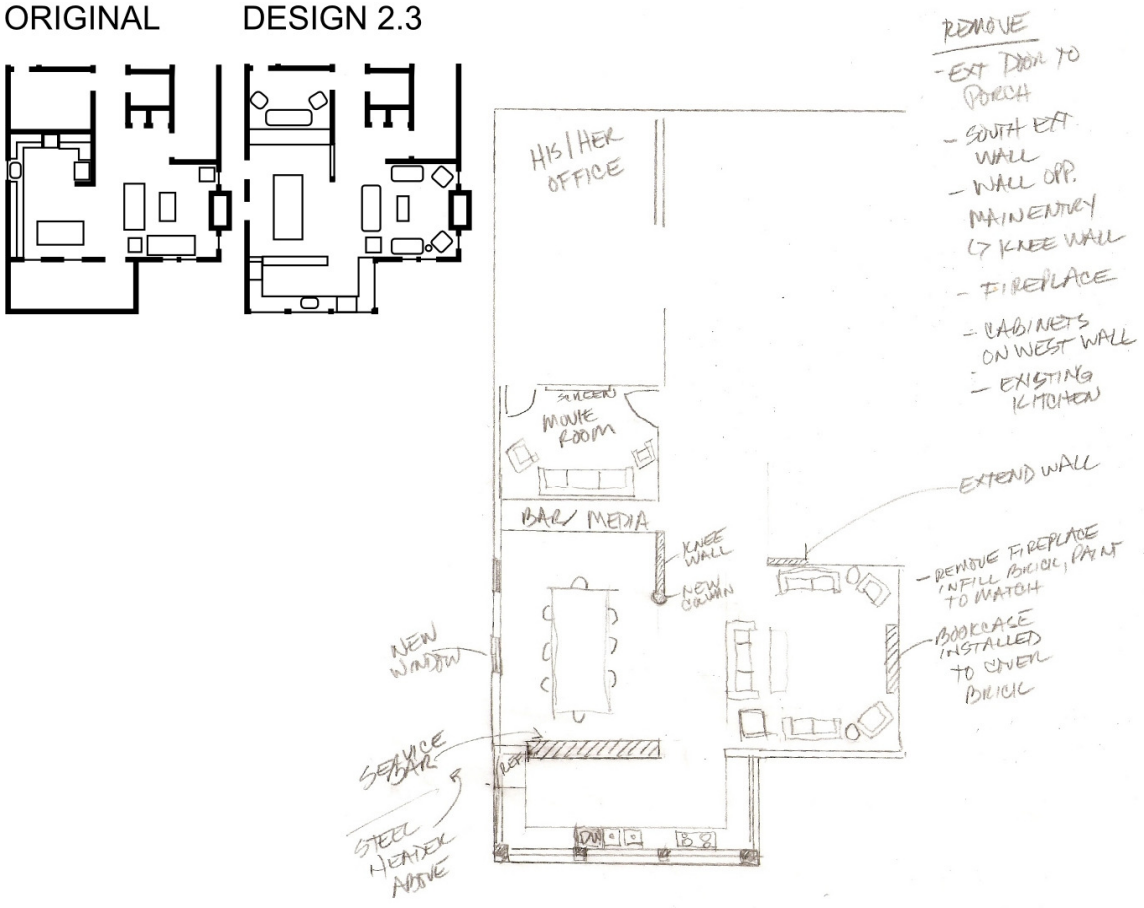


Figure 6.12: Design 2.3 title: "A space to bond." Architectural movements: integration of the balcony, segregation of the foyer, dedication of a media-centric space, visual linkage, space bounding. Architect moved kitchen into balcony, re-oriented dining room into kitchen, removed wall between kitchen and office, reprogrammed the office into a movie room, and extended the south wall of the foyer. Architect's statement: "Enhance the close relationship [between the clients]. Reprogram and combine some existing spaces. Recover unused space."

Figure 6.13: Design 2.4 title: "Life is 'CO-EXISTENCE'." Architectural movements: integration of the balcony, dedication of a media-centric space, visual linkage, space bounding. Architect removed north and west kitchen walls, integrated the balcony and the office, moved kitchen to east wall, created a media room in previous office space, and created a reprogrammable space. Architect's statement: "Even with technology, life's essence remains sharing the experience. Mobility. Boundlessness. Availability. Flexibility."

Figure 6.14: Design 2.5 title: "Framing study and view." Architectural movements: segregation of the foyer, visual linkage, space bounding. Architect created a window frame in west kitchen wall, extended the north living room wall with a window frame on it, and created a study on south end of the dining room. Architect's statement: "Open views from kitchen, entrance, dining room, and study bar."

Figure 6.15: Design 2.6 title: "explore the alternatives of programming the space." Architectural movements: integration of the balcony, segregation of the foyer, dedication of a media-centric space, visual linkage, space bounding. Architect moved kitchen into balcony, placed a large bookshelf, separating private from public spaces, removed the entrance door, and reprogrammed the office to be a media room. Architect's statement: "Enclose the private spaces. Open the public spaces."

Table 6.1: Summary of architectural movements for the 11 designs.

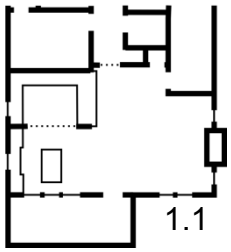
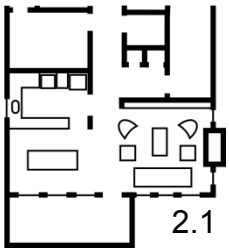
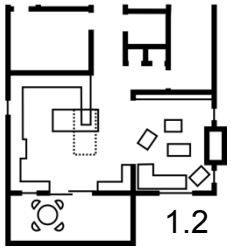
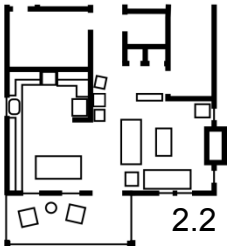
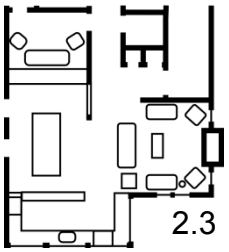
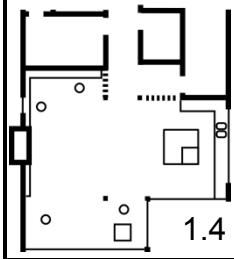
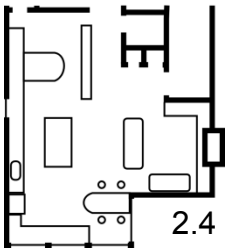
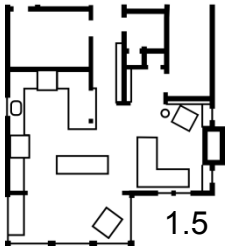
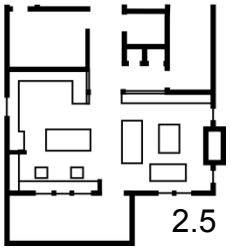
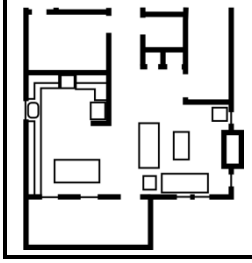
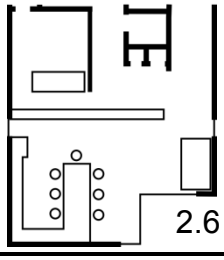
Group 1	Balcony	Foyer	Media Space	Visual Links	Bounding	Group 2 with Viz-A-Vis	Balcony	Foyer	Media Space	Visual Links	Bounding
 1.1				•	•	 2.1		•	•	•	•
 1.2		•		•	•	 2.2	•				
 1.3	•	•		•	•	 2.3	•	•	•	•	•
 1.4	•	•		•	•	 2.4	•		•	•	•
 1.5	•	•		•	•	 2.5		•		•	•
	ORIGINAL FLOOR PLAN					 2.6	•	•	•	•	•

Table 6.1 summarizes the comparative analysis of the designs of both groups. The main finding is that only the second generated a space dedicated for media consumption. Now, we move to the primary goal of this study, which is to collect and formalize evidence of Viz-A-Vis supporting behavioral pattern discovery. In the process of collecting and analyzing evidence, we came across a number of categories relevant to our discussion. We have pruned and grouped the categories and we have defined the relationship between them as they relate to our central hypothesis. Table 6.2 on page 207 contains a summary of the analytical structure. We classified the codes within the data into two categories: *CURRENT PRACTICES* and *DISCOVERY THROUGH VIZ-A-VIS*.

We synthesized CURRENT PRACTICES mainly from our observation of the first group. We are not claiming these observations can generalize to larger communities of practice. Simply, this is the synthesis of the sample of five senior architects. Yet, as in most expert analyses and participatory design studies, the discovery of new practices quickly reaches an asymptotic limit after a few participants (Seidman 1998). The analysis of this category gives us greater insight into the stages of design where Viz-A-Vis would be most useful. More importantly, it informs the practical methods of use of Viz-A-Vis within the architectural design process. Outlined, these are the stages of design: *Program Overview, Mental Sketch, Draft Behavior, Draft Sketch, Refine Behavior, Evaluate Sketch, Refine Sketches, Prepare Presentation, and Present*.

The Program Overview includes the presentation of the client, the design program, the design material, and the original clarifying questions. The design started with the presentation of the client, their stated requirements and the limitations for the renovation project. We distributed a written copy of the design program to each architect

and the client delivered the requirements in detail and with example instances of activity and behavior. We also distributed the material stipulated in the design program, namely, plans, elevations, and photographs. Then, architects asked clarifying questions, such as "in the floor plan, where exactly are the limits of the renovation?"

Next, we gave them a tour of the apartment. During the tour, the architects asked a number of questions to create a draft of the clients' behavior patterns. We classified the category Draft Behavior into two concept-classes: *Probing Questions* and *Testing Questions*.

Probing Questions help the architect create a draft of the client's environmental behavior patterns. We observed a number of types of Probing Questions. The most prominent matters of inquiry were working, cooking, playing, and entertaining. Architects asked about each one of these practices with mostly open questions.

"What do you do in the office?"

"What kind of cooking do you do?"

"Do you use the fire place?"

"How many guests do you usually invite and what do you do?"

Testing Questions help the architect determine the viability of design. The ratio of Probing Questions to Testing Questions in our study was five to one. Yet, Testing Questions revealed evidence of the state of design in the architect's mind: the Mental Sketch. There were two types of Testing Questions: *Expansive* and *Restrictive*. Expansive Questions were generic and tested the viability of large changes. For example, "would you like a larger view of the tree outside and of the Atlanta skyline on the background and what impact would that have on your privacy?" Restrictive Questions were specific

and tested the constraints to small changes. For example, “can I move the cameras that are on top of the television?”

Architects engaged in Expansive Questions evidenced, with greater clarity, the formulation of Mental Sketches. It appeared as though they had mentally formulated a rough sketch of their renovation during the clients’ exposition. By the time they were gathering information about the clients’ behavior, they were already engaged in a rapid hypothesis generate-and-test cycle. Restrictive Questions did not evidence a clear intention and seemed to be buying time for the architect, as the plans for their renovation formed in their minds. These questions did not seem to add to the design product.

Draft Behavior is the first stage in the current practices where we see utility for Viz-A-Vis. The goal of the questioning is to draw a rough sketch of the clients’ environmental behavior, their practices of flow and occupation of the space during different episodes of their lives. The exploration includes issues central to Viz-A-Vis: space, time, activities, and social context.

After the initial information gathering, the architects externalized and self-evaluated their propositions through Draft Sketches. We recognized two distinct methods of sketching: first, from the detailed object in perspective projection to the general plan in orthographic projection; and second, from the general plan to detailed objects. A number of our participants drafted multiple sketches and later evaluated their quality based on a number of metrics, such as aesthetics, functionality, and flexibility.

At that point, participants refined their understanding of the clients’ habits by asking more questions. They Refined Behavior. Most of these questions probed the lifestyle and behavioral patterns of the clients, for example, “do you like cooking?” and

"do you read or browse most of your books, or do they remain untouched?" The large changes to the environment had already occurred and the participants were working on the details of *programming the space* for multiple purposes through the re-use of places and objects. They used the phrase "programming the space" roughly to mean, "To create affordances within the space to support its assigned functionality." For example, they would think about the design possibilities for a sofa in the living room: watching television, reading, napping, entertaining guests, and working. At this stage, all the inquiries we recorded were Expansive Questions. They were not asking about the possibility of moving small objects or changing small spaces. By now, their designs had already taken direction. Rather, they were testing the viability of new programmatic alternatives to support the richness and variety of desired activity patterns. The focus was on the abstract behavioral patterns rather than the concrete spatial instantiation of those patterns. In other words, they were looking for ways to promote desired lifestyle by detaching it from the current, underperforming space, and re-attaching it to their proposed spaces. They abstracted behavioral patterns from current spaces and reified them into their proposed spaces. Refine Behavior is the second clear opportunity for Viz-A-Vis to grasp.

At this second information gathering stage, the architects did not ask for the clients' approval of a partial design, for instance, "do you like this re-distribution here or that furniture there?" Rather, they focused their inquiry into the behavior of the clients. The underlying attitude we observed in the architects' comments and gestures was that they did not need the clients to inform them about design. That was an emergent property

of the process. They did need, on the other hand, the clients to explain their preferred lifestyle in as much detail as possible and through multiple instances.

After gathering more information, our participants initiated a second pruning of their design space. They evaluated the quality of their design mainly by the degree to which the designs satisfied the requirements and fit the behavioral profile. This stage, Evaluate Design, is a possible third place for Viz-A-Vis to have impact. We reached this conclusion by the discussion on the focus group when our participants were stating the possible future applications of Viz-A-Vis. We will discuss this application in detail.

Finally, the architects refined their sketches, prepared the slides, text, and title for their presentations, and delivered them. We asked the architects to share their purpose of intent during the process and how the increase of information gathered from the clients influenced their train of thought. In their presentation, they had at least one "process" slide with a description of the design process. This slide, together with the verbalization of it and the observation of the design exercise, served as the basis for the description of the process we have given here.

From this analysis, it becomes clear that there are two stages in the process where Viz-A-Vis may make an impact. The first stage is in the original data gathering. The second, and more prominent stage, is in the refinement stage, where architects ask mainly behavioral questions from the clients. We presented Viz-A-Vis to the second group of architects at those stages: during the original description of the clients' behavior and during the refinement stage. From the two focus groups and from the design session of the second group we abstracted four broad concept-classes for the category DISCOVERY THROUGH VIZ-A-VIS: *Initial Concerns*, *Current Affordances*, *Possible*

Applications, and C*ritiques and Suggestions*. Each of these concept-classes parts into a number of sub-classes. Next, we will define and describe each. We will end our analysis tying these descriptions back to CURRENT DESIGN and to our original hypothesis for supporting discovery.

The category Initial Concerns groups together the questions and commentaries the architects had regarding the capabilities of the system. The reassuring questions generally emerged immediately after the presentation of the tool. For the first group, the questions surfaced at the beginning of the focus group. For the second group, they emerged at the beginning of the design session.

There were two types of Initial Concerns: C*apture Misses* and *Interpretation Issues*. First, Capture Misses encapsulates questions regarding the granularity, coverage, recall, and precision of the sensing infrastructure of Viz-A-Vis. These are sample capture miss questions:

"What is the logic of the position of the cameras?"

"What about putting cameras on the corners rather than overhead?"

"Would you miss my hand moving vertically?"

"How big are the pixels in the world?"

"Even if I don't move you, can still count my presence?"

Our participants were determining the sensitivity of the system to capture and process raw data. They understood the limits of the field of view of the overhead cameras, their resolution and frame rate, and the computation of motion through adjacent frame difference. Nevertheless, some doubt remained regarding certain details. The most prevalent concern was motion in the direction of a projection axis of the camera, that is,

directly to and from the camera. This question surfaced independently from three participants. To be clear, camera-radial movement generates concentric motion by adjacent frame difference. The motion is not lost. From our study perspective, this concern and its assurance are important because they depict the general category of Initial Concerns. Our participants wanted to be sure the data they were consuming represented reality with enough fidelity to allow them to arrive at sound conclusions.

Participants expressed a number of Interpretation Issues when consuming the data presented through the visualizations, namely, future behavior modeling, target behavior highlighting, person tracking, and color scaling. A recurrent concern for participants, both when seeing the data at the design session and the focus group was how to extract the behavior patterns from the current floor plan and project them to their proposed designs. A fruitful argument started around this issue. Some participants felt the tool did not give the mechanisms to project abstractions of behavior to new environments. Other participants disagreed and argued that detaching and reattaching behavior was what architects did when they interviewed clients. The participants defending the position of projecting behaviors viewed the practice as a fundamental skill of architects. “As an architect, do I care about the present use of space and is this tool a good starting point, giving a richer spatial signature of behavior than we had in the past?”

From the discussion, we raised the future possibility of adding behavior simulation to Viz-A-Vis. The idea we discussed in the focus group was to create a digital copy of the current environment with agents that model the current behavior of the clients and then place the agents on the new environment as a novel form of architectural design evaluation.

The second Interpretation Issue participants raised recurrently was target behavior highlighting. For example, participants wanted to distinguish visually between positive and negative instances of target behaviors. "When were the clients engaged in desired activities? You mentioned that you liked to be together, even if working in parallel you like the presence of the other person. But I can't distinguish those instances from every other behavior." Part of the problem with this issue is that the participants did not have long enough exposure to the mappings in order to gain enough experience to interpret them fluently. Each mapping we presented was an exercise of learning how to read it. By the end of the session, participants were fluent in mapping amount of motion to portions of space and detecting general patterns of behavior. Detailed descriptions of the mappings require longer exposure and access to the original files.

The third Interpretation Issue was person tracking. Four participants raised this issue independently. They wanted to be able to distinguish people in the visualization. Furthermore, they wanted to be able to query by person: "Can I see the action of just Rosa?" for example. The computer vision technology necessary for reliable person tracking under the constraint conditions of the Aware Home may be a reality today and it is certainly a direction to pursue in the near future. Person and object tracking is the natural next step. Implicitly, Viz-A-Vis is a space-tracking tool.

Finally, participants took issue with the color maps of the Activity Cube. The maps scaled over the entire cube, from dark blue to red. As a result, many layers within the cube only had shades of blue. The red zones were relatively scarce. The participants stated that they would have preferred a per-layer color map, even if it did not represent the data in exact fidelity. There are other coloring scales that may be used that would

have shown both the within layer variation and a faithful account across layers. Scale is a simple implementation feature that, nevertheless, raises serious issues with the visualization. This is a well-known fact in the visualization community. We mention it here because the participants raised it multiple times. In our original design, we had a slider for re-mapping the color scale, based on the intensity of target behaviors. Unfortunately, we did not implement it in our final design because of time constraints.

The concept-class Current Affordances is the central topic of our discussion. It is the synthesis of how users appropriated the analytical affordances of Viz-A-Vis. We group Current Affordances into the three representations of motion in Viz-A-Vis: the Activity Table (AT), the Activity Map (AM), and the Activity Cube (AC). Figure 6.16 shows the Activity Table, the Activity Map, and the Activity Cube displaying different representations and views of the same raw data. The data is the result to the query "What does typical cooking and eating look like?"

Participants described the Activity Table as "a linear representation of activities in space across time." They talked about its functionality. "From an architect's point of view, the AT can give a good summary on how space is used according to two dimensions: the intensity and the duration of activity." By "intensity," they meant the level of the aggregate at a particular block of cells in the table. For example, in Figure 6.16, zone 32, the north-west corner of the kitchen counter, contains intense activity during the first few minutes of the sequence. On the other hand, zone 23, the dining table, contains mellow activity for a relatively long period. The combination of these dimensions affords a discovery of patterns.

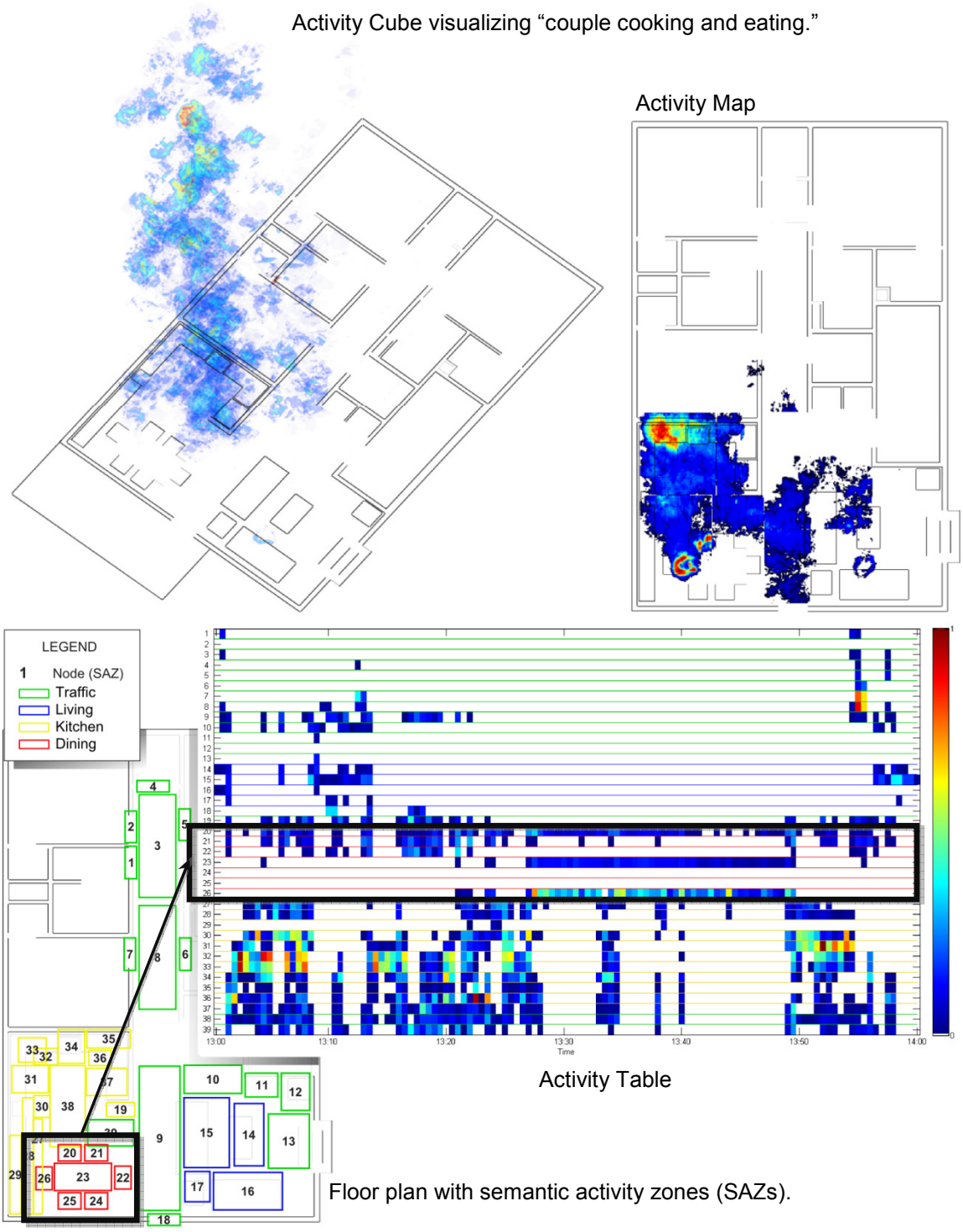


Figure 6.16: Results to the query: “what does a typical cooking and eating [activity] look like?” Top left, the Activity Cube. Top right, the Activity Map – 2D aggregate of all the layers of the cube. Bottom left, the semantic activity zones aggregating motion into places of interest. Bottom right, the Activity Table, mapping aggregate motion to in places to rows and time to columns. The color scale is constant across all representations. Notice the contrast between zones 20, 23, 26 and 24 and 25.

"[Looking at the Activity Table in Figure 6.16] The Activity Table can show us two kinds of information visually. It can show us how the use of space over time, so it's an agglomeration. It can tell us what kind of space is used and the way it is used. For instance, when you see the dining table, if I think about the red zone, it is something that is used much longer than the blue, which is the living room, the quality, uh intensity, of occupation is different. So, it can tell you, for instance, what is happening in the kitchen or the dining room is something that is really used pretty often but the living room is something that is very social and functional. That is not something that you think at first at your house when you have activities. You think you use a lot of the living room, but you realize you don't really use the living room much. So there is two kinds of information how long it's used and the intensity of actions. For example, the living room was used in a very condensed part of the overall time and the activity during that condensed period was very intense. The dining area on the contrary was used recurrently during the overall period of time recorded, but the activity was less intense."

Participants also highlighted the space-time properties of behavior. They generated a set of vocabulary words to describe the information AT visualized, among which are: "*distributed*, *punctual*, and *episodic*." By distributed, participants referred to activities that last long or span across large parts of the space. The reverse of that was punctual, where activities lasted short or occupied contained spaces. Figure 6.17 visualizes these concepts. There were other options, for example, distributed-short and punctual-long. The value of this exercise, for our evaluation, is that our participants engaged in an analytical process of abstraction. The tool afforded thinking about the

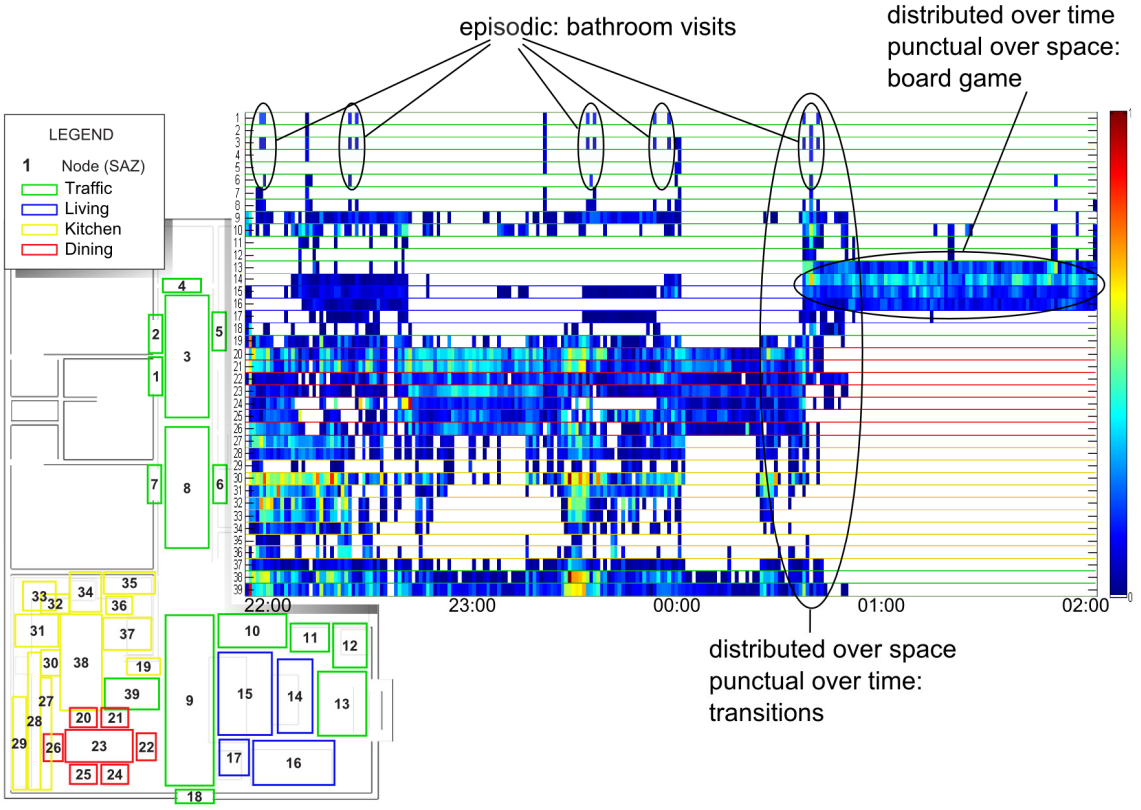


Figure 6.17: Activity Table with four hours of eight adults having dinner and playing cranium. Participants created a set of space-time categories to analyze behavior using novel vocabulary inspired by Viz-A-Vis: "distributed and/or punctual activities over space and/or time"

essence of activity in novel and abstract terms. Episodic activities had recurrent patterns of space utilization.

The Activity Map is the aggregate of motion over a single layer of space mapped on top of the floor plan. For many of our participants, this map was the most useful and easy to understand part of Viz-A-Vis. "This is a real measurement of what we predict with space syntax!" Space syntax is a theory and practice that analyzes the underlying structure of space (Hillier 1996). "An Isovist presents the *potentiality* of space. Isovists and other syntactic representations and metrics, describe the relatedness of space. Some activities naturally require that we position ourselves in space taking into account

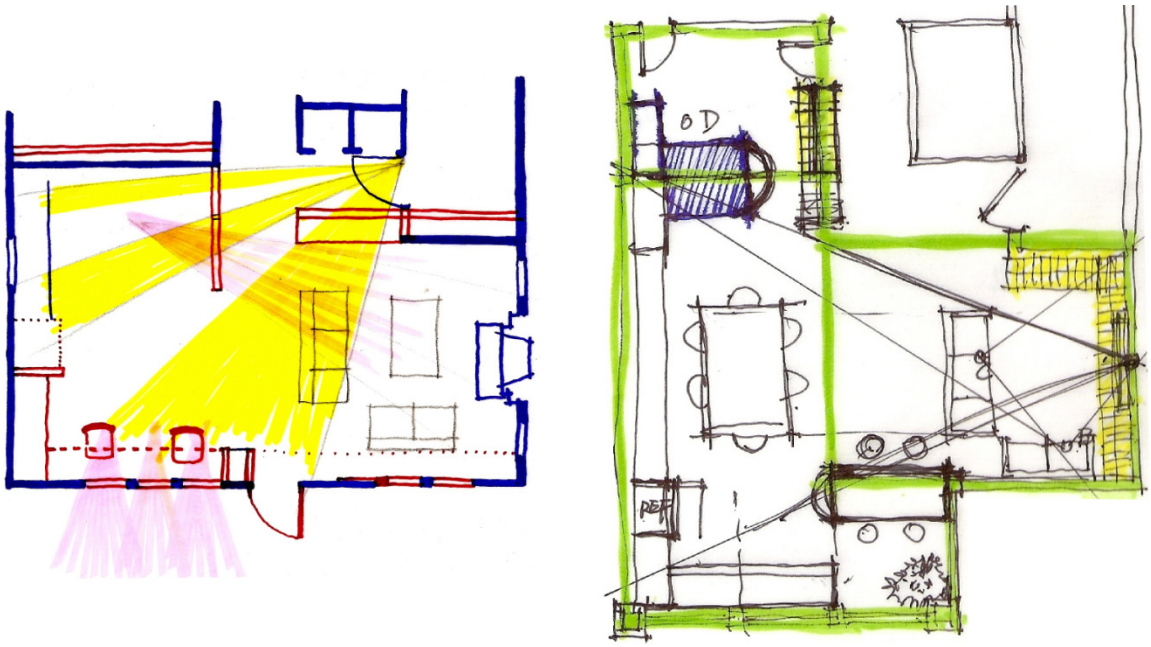


Figure 6.18: Two second-group participant sketches inspired by the Activity Map. Participants stated that they were "mapping the clients' desired patterns of behavior in terms of flow, movement, occupancy, visibility, and connectivity."

relatedness. When people concentrate on reading a book, for example, they may use an isolated corner of the living room, but when they want to read a paper while watching over the kids, they may sit on the most exposed chair. Syntax is about how spaces relate. Viz-A-Vis is about the intensity of activity in some locations, as well as the pattern of transitions between locations. "The Activity Map shows the *reality* of space." An Isovist is a computation of an inherent property of space. It discretizes a space under study and computes the visibility of each cell in the space from all other cells (Hillier 1996).

We observed two instances of design sketches directly influenced by the Activity Map. The authors of the sketches stated that they based the concept of the sketch on the visualization. Figure 6.18 shows the two drawings. The authors of these drawings gave similar explanations of the choice of diagramming. They intended to plan for connectivity, visibility, and flow based on the previous patterns of behavior they had

observed with the Activity Map. During the focus group, participants concluded that the Activity Map afforded conceptual diagramming. They saw the activities in the previous space and projected the desired behavior to the new environment. This discussion ties in with the discussion of Interpretation Issues, where the architects contended about the nature of requirement gathering and design instantiation. The participants who used the tool for diagramming used it to instantiate behavior. This affordance depends on personal practices more than most of the other features of Viz-A-Vis.

Most participants valued the simplicity of the Activity Map. They considered it the most obvious representation. Furthermore, they found many insights and discovered a number of patterns when we presented the episodic aggregates in Figure 6.19. The pattern that the focus group explored the most was "introversion." They correctly observed the pervasive lack of activity near the windows and balcony and concluded that the clients were introverted or, more precisely, "indoor-focused."

A point of debate was whether to design the space to support introversion, or to try to promote extroversion. Regardless, they all agreed with that observation, including the clients. The clients did not describe or considered themselves introverts, but when presented with the data, the discussion, and an exercise of introspection, they finally agreed that yes, on Tenth Street, they did not want to be by the windows. Tenth Street's publicness surpassed the clients' threshold of openness. In that environment, they were introverts. This was a discovery for the architects *and* the clients, the author and his wife.

The final point of discussion of Current Affordances is the Activity Cube. The participants found the cube to be the most challenging representation. It was hard for them to see where and when the activity patterns occurred. The cube suffers from the

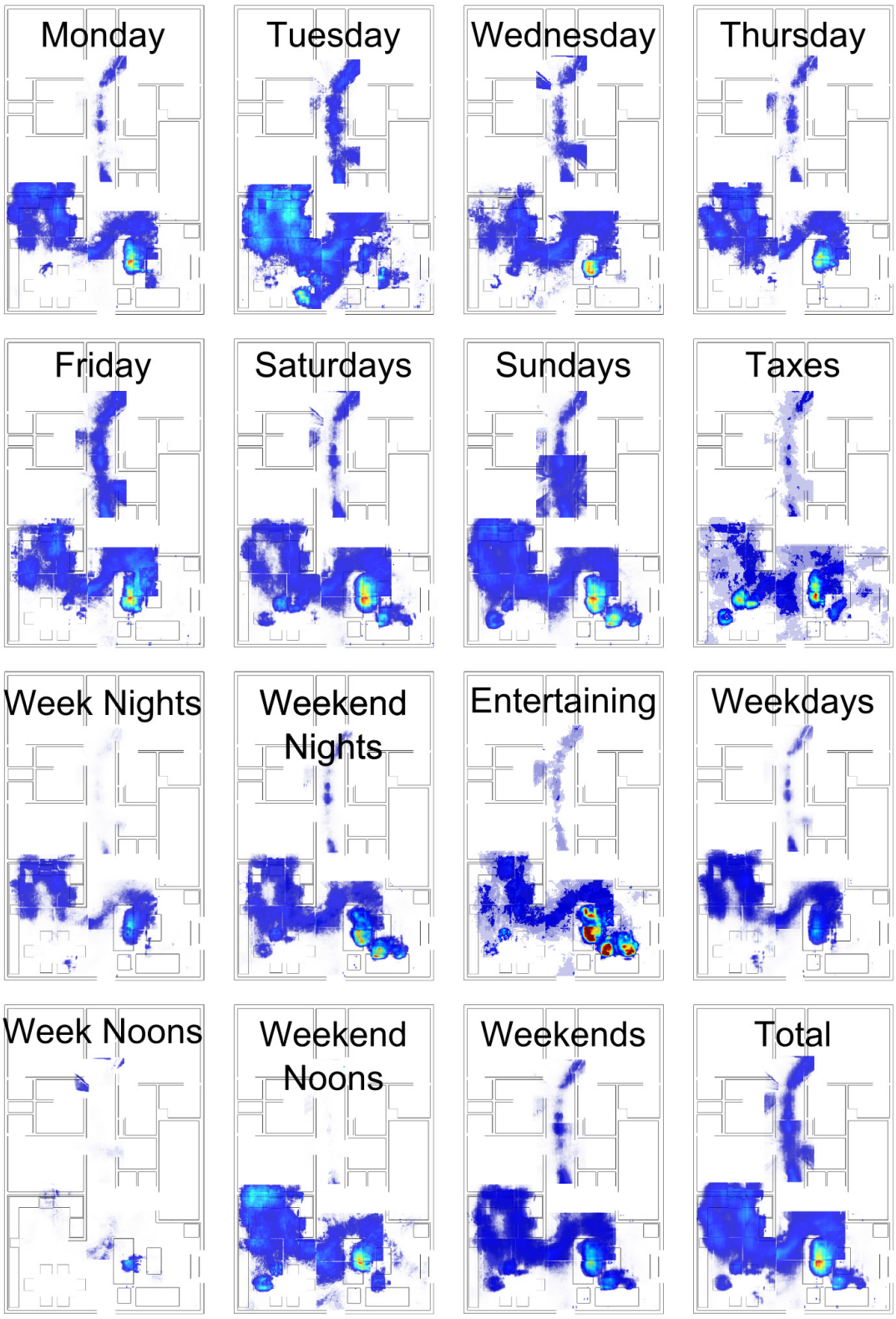


Figure 6.19: Episodic Activity Maps from the March 2006 9-day data collection experiment. Notice the behavioral patterns generally avoid the windows and the balcony, a sign of introversion or "center focus." This was a discovery for both the architects *and* the clients.

well know problems of 3D visualizations, namely, self-occlusion, depth ambiguity, and disorientation. One of the participants commented that he did not have enough time to digest the data; it "zoomed-by" too fast. Regardless, the cube prompted a number of engaging discussions. The most salient was concerning a perceived sedentary lifestyle of the male client. During the week of data capture, the client worked from home.

He worked mostly from the couch, zone 15. On Tuesday, he worked from the dining room table, (see Figure 6.19). On the Activity Maps, the aggregate of motion showed activity from the client. Upon further inspection, the architects discovered that most of the activity occurred during lunchtime. The Activity Cube afforded the segregation of the data and the visualization of finer temporal detail and sequentiality.

We designed the experiment such that architects functioned as high-cost professionals with very limited time. We supported all their non-design tasks, such as scanning and printing. We supported, as we have stated, the query process in Viz-A-Vis. They did not have to search for activity patterns. They only consumed pre-fabricated answers put together by the technician. It is in the process of searching for the patterns that the cube becomes most useful and where the most experience is necessary, both to manipulate the cube and to recognize the patterns. The architects did not go through that experience.

Participants stated a number of potential applications to the technology of this study. We created a concept-class, Possible Applications, and we grouped concepts into three types of Possible Applications; applications of: complexity, utility, and fidelity. This class allows us to study what the participants see in the tool but were not able to

exploit, or see as a natural next step. It is the continuation of the previous class, Current Affordances.

Applications of complexity mainly refer to the fact that the environment in which we tested and evaluated the tool is too simple for the architects to exploit the potential of the tool. This was not the opinion of all the architects, but enough raised the point that we include it here. All agreed that increasingly complex environments would exploit the tool to its potential, but not all agreed that the current environment and the verbal description of the clients was simple enough to understand everything about their behavioral patterns.

> "Because the relationship between the people occupying this space is not complex, I don't feel the data shows me anything that your description of the space did not show me already. In a more complicated program, for example, I have a building that is not working, and I don't want to sit and observe for 40 hours in a week, because there will be anomalies. And this could be in a hospital, a library, more complex environments."

> "This tool would help much more where you have organization and how to move to communicate, congestion, points of contact, like an ER or a traffic desk, or an urban area, a plaza, a congregate."

The second type of Possible Applications is applications of utility. The main observation of the participants was that the study had been brief and they had had little opportunity to explore the utility of the tool. Some participants formulated how they would incorporate Viz-A-Vis into their practices or the practices of other analysts.

> "Viz-A-Vis has potential to inform the design process at different stages and scales: object-design project – designing objects taking into account their place in

the environment; the small-scale design process (housing, everyday space); the larger project (airport, crowd) – individual pattern, social patterns, crowds patterns. However, the visualization of the data and the type of data recorded should be adjusted accordingly."

"We can approach it from a functionalist perspective: use this tool to rearrange the spaces to improve the traffic between them and in them."

"The study could focus on smaller areas rather than larger spaces that architects normally focus on could bring up things that architects were previously unaware of."

"The tool would be more effective if the home could be shifted since there is only one "answer" to most movement issues."

"Design a vacation house – use the data from a family to modify new environments to fit the needs of the specific family when they vacation in a home that is not their own. The flow matches their natural rhythms."

"Design a museum space where the curator distinguished patterns of *translation* versus patterns of *vibration.*"

This architect was referring to the natural behavioral tendencies incurred during Art exhibitions to move-on versus contemplate. From the curator's perspective, it is important to be aware of which space programs produce opportunities for contemplation, reflections, introspection, and dialog. From this perspective, the focus group discussed the two types of movement we had encountered in our analysis of motion with Tableau Machine (chapter 3). We discussed vibration, the type of motion that does not produce a change in place, and translation, the type of motion that produces changes in place.

Recall, place is socially or programmatically defined space. Typical vibrations include hand and head gestures and fidgeting. The system captures and aggregates this type of motion. Paradoxically, this type of motion always produces greater aggregates than translation, a type of motion that generates a change in place. The reason for this phenomenon is that, unlike translations, vibrations occur constantly. When treating the concept-class Critiques and Suggestions, we will go deeper into the issue of differentiating abstract types of motion.

Other types of utility applications include analyzing shopping patterns at large department stores, generating topographies of movement similar to the patterns dance and sport coaches strive to generate, but from a different perspective.

The final class of Possible Applications is applications of fidelity. Here, participants are referring to the natural noise of collecting data, both objective and subjective. They saw the capture infrastructure of Viz-A-Vis as an opportunity to collect data automatically and with fidelity.

> "When you sit and observe, you learn that sometimes what people say isn't what's going. For example, you said the dining room is extremely important to you. The data showed me that, in fact, the kitchen and the living room are more occupied than the dining room. That tells me one of two things: (1) your current configuration is not working and you would want the dining room to be the center of your activities; or (2) you have an ideal of what you want that is not really what you want, and a new design based on that ideal will have no positive impact. With you tool I can compare what you are hearing with what really happens."

The final concept-class is Critiques and Suggestions. Here, participants are not projecting towards a possible future, but promoting changes to current design choices. Critiques and Suggestions has three sub-classes: activity differentiation, people differentiation, and usability.

We have already mentioned activity differentiation in the discussion of Possible Applications. The activities participants were differentiating were vibration and translation. The architects shared enough opinions and ideas regarding this topic that we created this class for it. Other types of motion include "ritualistic motion" and "desired behavior pattern." We contained this discussion to computable types of motion. A simple algorithm should classify most vibration by comparing current displacement to a threshold-of-displacement. Again, we are looking for abstract classes to define their computation with flexibility and reliability. We are not looking for the solution to activity recognition.

The most abundant critique we received was the lack of people differentiation. We independently received the same question four times.

> "Can you tell the number and identity of people in the same space?"
>
> "It might be more effective if you show trajectories of travel. With the current method, if you visualize a plaza, it would be difficult to tell who was moving where over long periods."
>
> "It's not about individuality [the current implementation]. It's just a movement over space. The narrative you are missing is the people. Who generated the motion? I feel that the personal properties will have an effect that is not present now."

"Is it simple to output people – how they use space – aggregate according to how long they use the space."

The final type of critique we received was about usability. People wanted to address the tools ability to accommodate individual working patterns.

Table 6.2: Summary of the analytical structure of the focused coding.

- CURRENT PRACTICES
  - Program Overview
    - Design Program
    - Client Description
    - Clarifying Questions
  - Mental Sketch
  - Draft Behavior
    - Probing Questions
    - Testing Questions
      - Expansive
      - Restrictive
  - Draft Sketch
  - Refine Behavior
    - Probing Questions
    - Expansive Questions
  - Refine Sketch
  - Prepare Presentation
  - Present
- DISCOVERY THROUGH VIZ-A-VIS
  - Initial Concerns
    - Capture Misses
    - Interpretation Issues
  - Current Affordances
    - Activity Table
    - Activity Map
    - Activity Cube
  - Possible Applications
    - Complexity
    - Utility
    - Fidelity
  - Critiques and Suggestions
    - Activity Differentiation
    - People Differentiation
    - Usability

6.5 Discussion

The author played four roles during the study. He played the creator of Viz-A-Vis, the evaluator, the client, and the technician. Although all were real versions of him, we made it very clear when each was interacting with the participants. The main reason for revealing the authorship of Viz-A-Vis was to motivate participants by offering a real opportunity to have an impact on the tool. As the creator of Viz-A-Vis, the author strived to give detailed and faithful accounts of its workings and limitations. He did not explain how the interface works and he limited the description of the affordances of the tool to one simple example: "this is what arriving looks like." As the evaluator, the author strived to avoid leading in any architectural design-related decisions. He performed open questions both during the design sessions and, specially, during the focus group. During the design session he observed, took notes and photographs. During the focus group, he guided the discussion. As the client, he prepared the descriptions of requirements, answered questions about behavior, activity, lifestyle, and avoided leading the design, stating preferences only when asked. As the technician, he took careful notes on the queries and verified his interpretation beyond a doubt, performing queries, presenting results both in aggregate and original format. He carefully avoided interpreting the results in any architectural or design perspectives. As the technician, he simply stated the facts without interpretation.

The author carefully rehearsed all roles before the start of the study. Nevertheless, we consider this to be a point of contention. With greater resources, one or more people would play each role. The designer would be a different person from the evaluator and, most of all, from the subject of observation of the participants of the study. Having stated

that, we gathered very rich data, despite all the obstacles posed by this design. Next, we will discuss the issues in detail.

An issue that neither participants nor researchers brought up during the focus group discussions was the fact that the fictional clients were extremely aware of their self-behavior. First, they were the subjects of the intensive data gathering session of the March 2006 experiment. Second, they kept a detailed journal of the experience. Third, they transcribed and described parts of the overhead camera data. Fourth, they carefully studied their behavior in order to design this study. Fifth, an expert architect skillfully advised the author in the formulation of the fictional clients. Sixth, the players for the clients rehearsed the descriptions of their behaviors and answers to possible questions in order to report the same data to both design groups. Very few clients will experience this level of familiarity with their own behavioral patterns. The logical implication to this study is that we severely limited the opportunity for discovery with the capture and visualization tool by providing an extremely rich and faithful account of self-behavior. Nevertheless, both participants and researchers were cheerfully surprised by the discovery and evidence-based argumentation of unexpected behavioral patterns.

Another point of contention is that by playing both the role of subject of observation and observer, we may have limited the freedom with which our participants shared their insights. We may have lost the data from cases where the participants discovered patterns that they considered are not prudent to discuss. We were aware of this problem. We created a very informal and amicable environment where they could feel greater freedom in sharing their observations, even if critical or negative.

Furthermore, we tried to leave some ambiguity on whether the fictional client was exactly like the researcher and his wife, or just based on them.

The third point of contention is the fact that the evaluator and the designer of the tool were the same person. Again, participants may have been less critical in the presence of the designer. Yet again, we promoted an amicable environment were critiques, no matter how harsh, were fully welcomed and required. Part of the rationale behind exposing the designer was to encourage participants to be part of the study with the potential reward of having direct impact on the final design of Viz-A-Vis.

Another problem for the evaluation of Viz-A-Vis with a hypothesis driven approach is that we could not perform open coding, and thus, Grounded Theory analysis. In other words, our analysis was contained by what we were looking for, thus, potentially, limiting possibilities for an emergent theory of discovery. Our research was not data driven. It was hypothesis driven; thus our choice for focused coding over Grounded Theory.

Our initial study design included an evaluation of the design product by an external panel of distinguished judges. Early in the study we recognized that the internal confounding factors greatly outweighed the external independent variable. Due to their scheduling limitation and by pure chance, the two groups were heavily unbalanced. One group had substantially more experience designing. At that point, we eliminated the external panel competition and re-focused the study on the verbalization of participants.

The goal of first focus group was to re-visit the designs based on the new information provided by the visualization. Participants could comment on their original designs or go as far as propose changes. Their response was that they did not need to

change their designs because the data was evidencing the same behavior as the verbal delivery of the clients. We had prepared for that event. We continued the discussion with a focus group, were the participants or the first design group shared their understanding and insight into how and in what context they would use Viz-A-Vis. The participants of the second design group had had experience consuming Viz-A-Vis visualizations. Their focus group had different emphases, as we've described earlier.

6.6 Conclusions & Contributions

We have given direct evidence supporting this study's research question. Viz-A-Vis supported a group of domain experts in discovery tasks central to their practice that would not have emerged without the tool. Furthermore, participants generated new concepts, language, and cognitive structures to support the categorical analysis and discovery of activity patterns. Specifically, the group of architects discovered behavioral patterns in their subjects of study, the clients, which the subjects themselves were not aware of. After careful consideration and revision of the evidence, architects and clients agreed on the validity and value of the discoveries. The most impactful was the discovery of the introverted patterns of behavior of the clients.

We have also detected a pattern of architectural movements between conditions. Four out of six participants who received exposure to Viz-A-Vis included media-centric spaces in their architectural movements. Not a single participant from the first group designed a dedicated media-centric space in their architectural movements.

We present the conclusions to this document and collect the directions for future work in the next and final chapter.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

This dissertation has presented the design and evaluation of two overhead camera visualization systems: Viz-A-Vis and Tableau Machine. We have aimed to create practical tools and artistic artifacts that support objective analysis and creative interpretation of activity in natural settings.

To accomplish these goals, we proposed the thesis statement:

In the process of overhead video interpretation and analysis of activity, combining computer vision abstractions with information visualization techniques provides: (1) improved user task performance measured by time to task completion, precision, recall, coverage, and user assessment; (2) improved user experience measured by user preference; (3) increased user capacity to discover activity patterns; and (4) new opportunities for creative interpretation, experimentation, conversation, and reflection regarding everyday activities.

We tested these statements through three empirical summative studies. We assessed statements one and two through a task-centric user study measuring performance and preference. We evaluated statement three through an ecologically valid domain-expert study. We tested the fourth statement with a longitudinal, in-situ qualitative user study. We answered our research questions with success and, more importantly, we opened new and exciting opportunities for future work.

Can computer vision abstractions and information visualization techniques improve the interface to analyzing activity in overhead video as measured by time to task completion, precision, recall, coverage, and user assessment (Thesis Claim 1)? Yes. With statistical significance, Viz-A-Vis outperformed standard video playback five-to-one and the state of the art video cube two-to-one in the task of searching brief, sporadic, unpredictable, and isolated events. Moreover, Viz-A-Vis outperformed standard video playback two-to-one in bounding long events.

Can computer vision abstractions and information visualization techniques improve the user experience of activity video-analysis as measured by user preference (Thesis Claim 2)? Yes. Users cited searching, overviewing, and discovering isolated patterns of activity as the primary tasks where they prefer Viz-A-Vis. Furthermore, most users found the Activity Cube infrastructure to be more engaging and to open interesting possibilities for the discovery of patterns.

Can vision-based data abstractions improve the information visualization interface as measured by analytical discovery of activity patterns (Thesis Claim 3)? Yes. Viz-A-Vis supported a group of domain experts in discovery tasks central to their practice that would not have emerged without the tool. Specifically, the group of architects participating in our study discovered behavioral patterns in their subjects of study, the clients, which the subjects themselves were not aware of. After careful consideration and revision of the evidence, architects and clients agreed on the validity and value of the discoveries. We also detected an effect of Viz-A-Vis in architectural movements between conditions. Four out of six participants who received exposure to Viz-A-Vis included media-centric spaces in their architectural movements. Not a single participant from the

other group designed a dedicated media-centric space in their architectural movements. Finally, and most impressive, the group of architects dynamically generated a new set of concepts and terms to describe, categorize, and analyze activity patterns. Pushed to its logical conclusion, this process of methodical analysis would lead to the creation of new theories and practices regarding environmental psychology.

Can a vision-based visualizing Art installation engage users in a long-term process of creative interpretation, experimentation, conversation, and reflection (Thesis Claim 4)? Yes. Tableau Machine succeeded in being engaging over a period of eight weeks. The incorporation of Tableau Machine into family life provided powerful evidence that its success is not purely a function of humans being able to read meaning into almost anything (a Rorschach effect), but rather that Tableau Machine's active interpretation and generation supported human playful ongoing experimentation and creative interpretation.

In this dissertation we have taken an important step towards supporting human analysis of activity captured through overhead video. We have positively argued for the intrinsic value of continuously capturing activity with overhead video. Moreover, we have demonstrated the practicality of the approach by computing a number of abstractions to visualize the relevant content hidden in the vast video collections. Furthermore, we have proven the intrinsic discovery affordances of visualizing the hidden spatiotemporal structure of human activity in its natural place and period. Finally, we have opened a vast research agenda for potential applications of visualization of activity through vision, Viz-A-Vis.

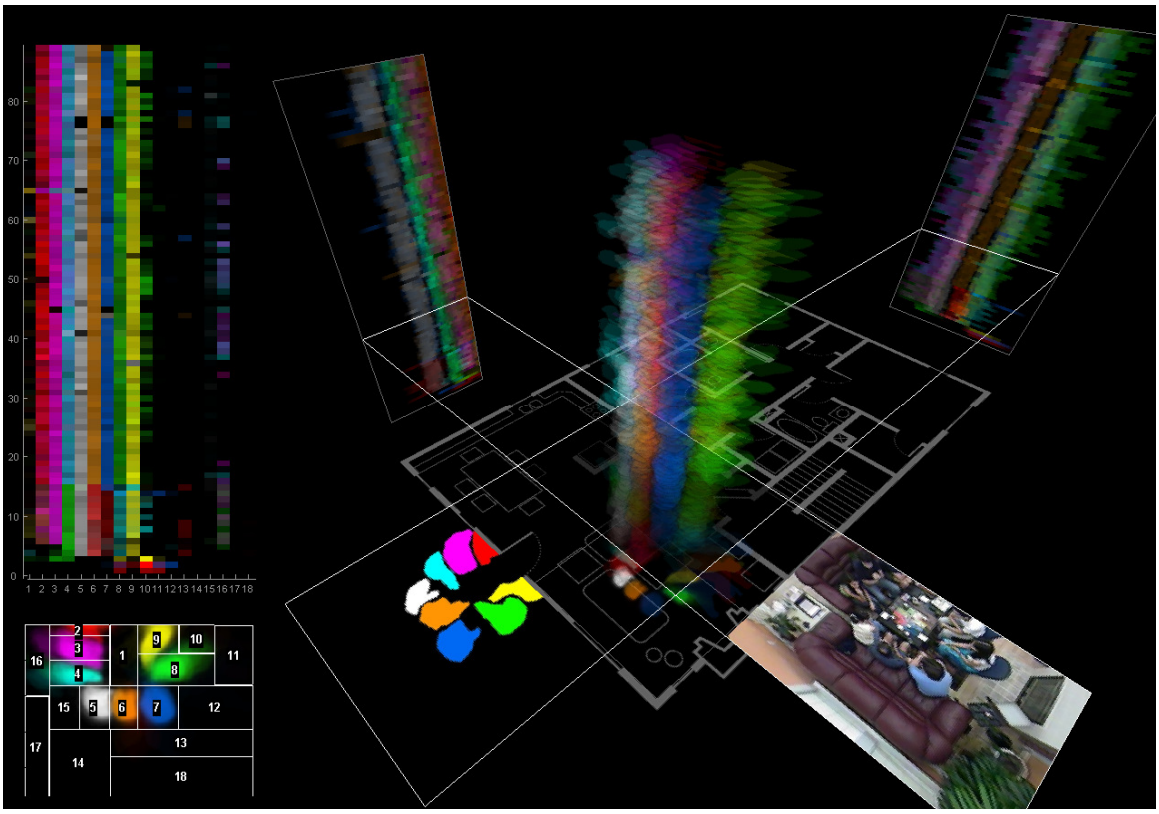


Figure 7.1: Blob-based visualization for tracking and filtering identity.

7.2 Suggestions for Future Research

The logical next step for Viz-A-Vis is to contribute in real environments that measure and analyze behavior. Our goal is to introduce the tool into real practices where established goals, deadlines, validation criteria, and procedures sharpen the tool. There is a great opportunity to develop Viz-A-Vis further. The key to future developments, in our view, is to drive them by real needs. For example, the most common complaint users in both studies had was the loss of identity. Architects highlighted that the lack of identity in the visualization created a hole in the "narrative of space." They could not answer "who" questions. In the user performance test, the main reason users cited for ranking Viz-A-Vis extremely low in the tasks of description and tracking was the lack of identity in the cube. Figure 7.1 shows a Matlab prototype of Viz-A-Vis including blob information. Blob

tracking is significantly more complex than computing and aggregating motion. Thus the benefits must fully justify the costs of each improvement.

There is a practical alternative to blob tracking. It involves combining the video cube and the Activity Cube. Put briefly, aggregate motion maps to the translucency layer of original video frames. The result shows only the relevant pixels of the video cube, where relevancy is a function of aggregate motion. The key is to provide the ability to view details directly from the "new" activity cube.

Another key area of immediate improvement is the definition of semantic activity zones. Currently, they are manually and statically defined. There is opportunity for growth in three areas. First, users should interactively define zones of any shape. Second, the machine should define zones dynamically, based on the statistics of space usage or the tracking of furniture and other large objects. Third, a mixed-initiative approach should define zones dynamically and interactively.

A third area of immediate improvement is the addition of interactive filters to Viz-A-Vis. Currently, Viz-A-Vis filters activity by space and time only. There are clear opportunities to filter activity by amount and type of motion. Currently, many fast activities such as translations are lost by the continuous aggregation of slow but steady motion typically generated by vibrations. In order to be able to study different types of activities, the user needs to be able to segregate by levels of aggregate motion and by displacement of motion.

A fourth immediate improvement is adding more interaction to the Activity Table. Currently, AT shows a static linearization of two-dimensional space. The resulting

visualization does not sustain spatial adjacency. The table needs to be able to hide and re-order rows to sustain local adjacency and provide visual focus to the analytical tasks.

Fifth, Viz-A-Vis needs to integrate the computation and aggregation of motion to the visualization interface. Online aggregation of thousands of hours of motion is not an efficient procedure. Some techniques may shed some light, for example, wavelet representation of partially aggregated video may be a practical optimization for the problem of semantically and temporally zooming with Viz-A-Vis.

Sixth, in order to promote a simple interface, we will include keyboard navigation to the Google Sketchup implementation of Viz-A-Vis, completely subsuming the video player. Furthermore, additional cognitive support structures need to provide scaffolds to higher-level inductive analysis, where the observations gathered through Viz-A-Vis group to form concepts, categories, classes, and theories of activity.

Finally, in the domain of visualizing computer vision, Viz-A-Vis is one of the first published attempts that explicitly explores the questions of mutually augmenting both fields. There is great room for improvement in the mixed-initiative interaction model that includes computer vision and information visualization to narrow the semantic gap between raw video and results. A number of techniques from machine learning and pattern recognition may find their way into working solutions. At first, progress will typically be that of applying known techniques to new problems. There is, however, rich ground for cross-foundational work.

APPENDIX A

FORMAL DEFINITION OF SOCIAL ENERGY, DENSITY AND FLOW

Motion

For all the *i* and *j* pixels, the difference image *d* at instant *t* is defined as:

$$d_t(i,j) = \begin{cases} 0 & if \quad |f_t(i,j) - f_{t-1}(i,j)| \le \varepsilon \\ 1 & otherwise \end{cases},$$

where ε is a positive threshold and $f_t(i,j)$ is the i^{th}, j^{th} pixel of the frame *f* at time *t*. When the value of $d_t(i,j)$ is *1*, we refer to it as *active pixel*.

Activity

We define activity over a SAZ at time *t* ($a_{SAZ,t}$), as the sum of active pixels weighted by the surface area of the zone,

$$a_{SAZ,t} = S_{SAZ} \sum_{\forall i,j \in SAZ} d_t(i,j),$$

where S_{SAZ} is the ratio of the physical area over the image area of the SAZ.

Social Energy

Formally, we define *Social Energy E* at place *P* as the amount of activity accumulated over groups of semantic activity zones over a window of time.

$$E_P = \sum_{\forall SAZ \in P, \forall t \in T} a_{SAZ,t},$$

where *t* is an instant in time, and *T* is the size of the temporal window. Both *P* and *T* are defined by observing the activity in an environment. In terms of the topology, *P* is a sub-graph composed of the SAZs that belong to the same contextual everyday use.

Social Density

In physics, Density is defined as the ratio of the mass of a substance to its volume. During the coding of home activity, we were metaphorically inspired by physical concepts to define axial categories. For example, we define Social Density as the ratio of social mass to its volume.

In physics, mass is defined as the quantity of matter in a body regardless of its volume or of any forces acting on it. Our equivalent of physical matter is social mass. We define *social mass* as the number of active nodes, regardless of the Social Energy or Social Flow of the activity. We define *active nodes* ($A_{SAZ,t}$) at time t as the SAZs that have activity above a threshold θ:

$$A_{SAZ,t} = \begin{cases} 1 & if \quad a_{SAZ,t} \geq \theta_{SAZ} \\ 0 & otherwise \end{cases},$$

where θ_{SAZ} is a per-zone threshold we learn from historic data as a fraction of the activity mean over a large window of time (several hours), i.e., $\theta = \alpha a'$ with $0<\alpha<1$.

We define *social mass* at place P and time t ($M_{P,t}$), as the aggregate of active nodes in P:

$$M_{P,t} = \sum_{\forall SAZ \in P} A_{SAZ,t} .$$

In physics, volume is a quantification of how much space an object occupies. Space in classical physics is three-dimensional. We abstract the physical and image space into topological space. In the topological space, volume is a one-dimensional quantity. The space between two nodes is the sum of the edges of the shortest path between the

two nodes. Recall that in our topological model the edges are weighted by the physical distance between the centers of the SAZ's.

We define social volume at instant t as the length of the minimum weighted spanning tree (*MWST*) that connects all the active nodes $A_{SAZ,t}$ of the sub-graph P of the adjacency graph AG:

$$V_{P,t} = length\,(MWST\,(AG, A_{SAZ,t})) \quad \forall SAZ \in P\,.$$

We have defined instantaneous mass and volume. These quantities are very dynamic over time. Instantaneous Social Density is social mass over social volume at time t:

$$D_{P,t} = \frac{M_{P,t}}{V_{P,t}}\,.$$

We define Social Density as the sum of instantaneous densities over a temporal window T:

$$D_P = \sum_{\forall t \in T} D_{P,t}\,.$$

Notice that Social Density is a discrete measurement. The topology is discrete, thus mass and volume are discrete.

Social Flow

Everywhere in our model, the spatial and temporal resolutions have been finite. If they were infinite, aggregate social dynamics would be computed as derivatives and integrals. Given the finite resolution of the sensors, we restrict our definitions to be sums and differences. For Social Flow we decided to leave the definition in terms of partial

differentials because we deemed the notation more straightforward than partial differences. In practice, though, they are only differences.

We define Social Flow as the transfer of activity between adjacent SAZ's. We measure the transfer of activity as the disappearance of activity in one zone and the appearance of activity in an adjacent zone. We quantify the rate of change of activity as the partial differential of activity in a zone with respect to time. Therefore, increasing activity is defined as a positive derivative and decreasing as a negative derivative. Formally, we define the instantaneous Social Flow between adjacent semantic activity zones Q and R at time t as:

$$F_{Q,R,t} = \begin{cases} \dfrac{\partial a_{R,t}}{\partial t} & \text{if } \dfrac{\partial a_{Q,t}}{\partial t} < 0 \text{ and } \dfrac{\partial a_{R,t}}{\partial t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

This definition creates a flow matrix for all the edges in the adjacency graph. Flow needs to be aggregated both in topological space and time. Thus, we define Flow F over a place P as:

$$F_P = \sum_{\forall (Q,R) \in P, \forall t \in T} F_{Q,R,t} .$$

Note that we define instantaneous Flow in terms of the activity of the incoming node. Social Flow can be caused by a change of position in physical space, but the two are not equivalent. Other actions can cause Flow. For example, people taking turns in talking and gesturing will activate a Flow between the SAZs they occupy because there is a decrease in activity in one zone and an increase in the other, assuming the zones are adjacent. This model does not capture Social Flow between non-adjacent zones. For example, two people gesturing to each other over long distances. Figure A.1 shows a three-way scatter plot of EDF in the five regions of the home.

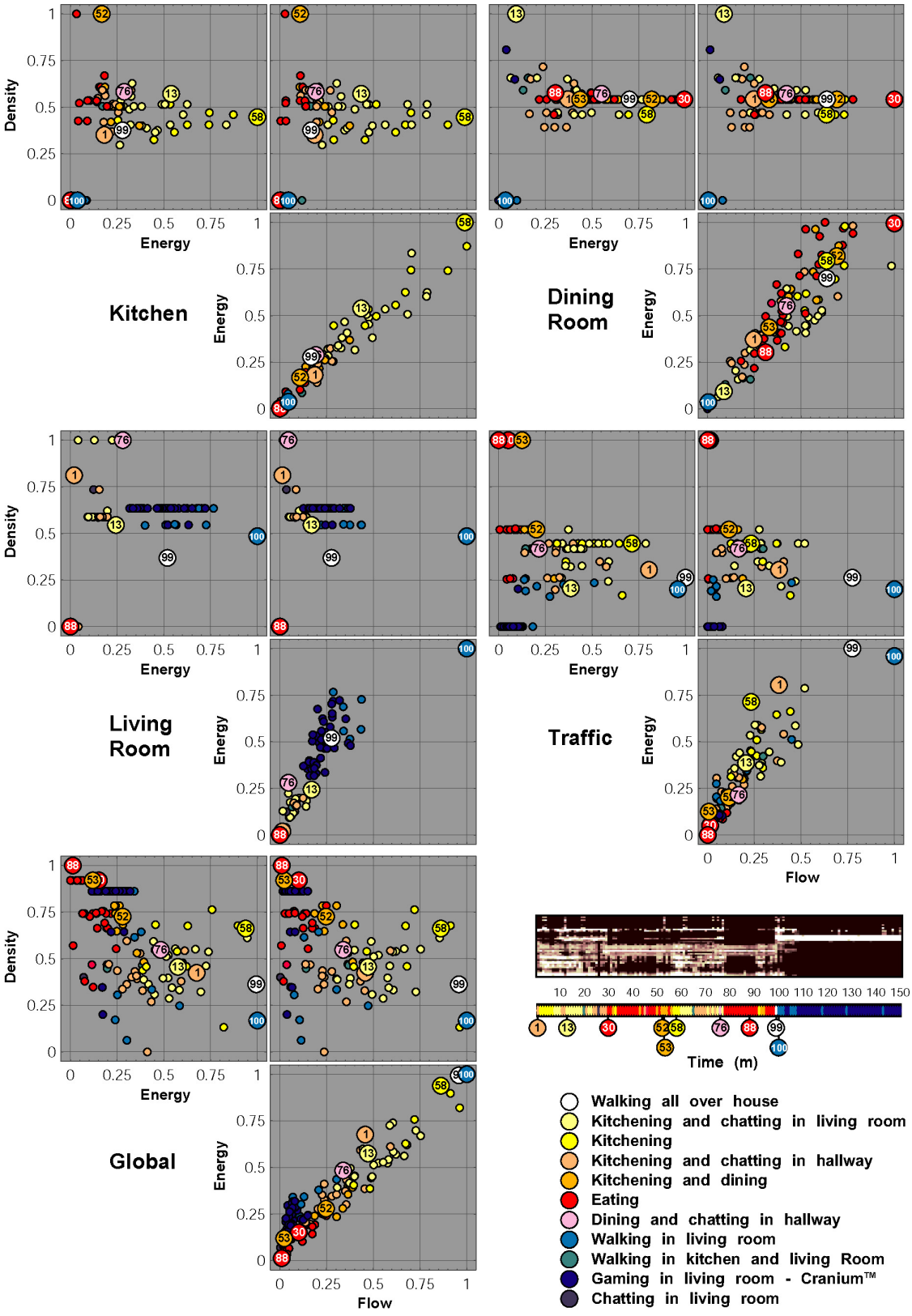


Figure A.1: Scatter plot matrix of Energy vs. Density vs. Flow for the Kitchen, Dining Room, Living Room, Traffic, and Global nodes.

APPENDIX B

GLOSSARY

Accountability (p.5): a system's account of its actions allowing the user to understand and recover from errors.

Activity Characterization (p. 78): a rapid data transformation from raw sensing to abstract and ambiguous representations of activity.

Activity Cube (p. 118): a three-dimensional view of aggregate motion mapping time to height.

Activity Map (p. 119): a floor plan view of aggregate motion.

Activity Table (p. 37): a tabular visualization of aggregate motion across space mapped to rows and time mapped to columns.

Adjacent Frame Difference (p. 89): computing motion in image sequences by differencing frames across time.

Bounding (p. 125): the user task of finding the beginning and/or end of a long lasting event, like dinner.

Co-interpretation (p. 28): a creative cycle of meaning making between an author and an audience actively mediated by an expresive AI system.

Counting (p. 127): the user task of enumerating the occurrence of short and repetitive events.

Coverage (p.): the length of video traversed during the task.

Cranium(TM) (p. 123): a board game that includes physical activity.

Creative Interpretation (p. 2): a subjective process of negotiated meaning making.

Describing (p. 124): the user task of translating video into English accounts of activity including actors, places, objects, and interactions.

Dominant Reading (p. 29): the understanding of an encoding where the author explicitly states the meaning of the message, leaving little room for interpretation.

Expressive AI (p. 22): a symbiotic practice of exploring new media through Human-AI interaction.

Interpretative Affordance (p. 29): a systemic feature that allows the audience of an artifact to negotiate the meaning of the artifact with its creator and, in the case of artificially intelligent artifacts, with the artifact itself.

Interpretative Scaffolds (p.29): a systemic feature of an artifact that engages the audience in dominant readings with the purpose of enticing the audience toward the more complex negotiated readings.

Mixed Initiative Computing (p.14): a symbiotic human-computer system that solves complex problems by combining the strengths and limiting the weaknesses of the computer and the human. The computer is good for crunching large numbers and bad for creative insight and viceversa.

Motion (p. 89): luminance change in image sequences assumed to be the product of physical motion.

Negotiating Meaning (p. 29): a balance between dominant readings and illegibility that affords creative interpretation.

Overhead Video (p.33): image sequence recorded from a static camera on the ceiling facing directly down that afford one-to-one mappings between pixels and locations.

Place (p. 92): socially defined space.

Precision (p. 131): the percentage of correct instances from the set of retrieved instances.

Recall (p. 131): the percentage of retrieved instances from the set of target instances in the original video.

Raclette (p. 123): an electric grill for the dining room table.

Reification (p. 95): the opposite process of abstraction, where an abstract representation is made concrete.

Rejected Reading (p. 29): a complex encoding indistinguishable from randomness that the interepreter rejects as meaningless.

Searching (p. 127): the user task of finding the occurrence of a short, sporadic, sparse, and unpredictable event, like a bathroom visit.

Semantic Acitivy Zone (p. 35): a place of everyday action.

Semantic Gap (p. 12): the difference between two representations with varying degree of abstraction of the same concept.

Social Density (p. 36): the inverse distance between active places at a given moment.

Social Energy (p. 35): the ammount of activity at a given place and moment.

Social Flow (p. 36): the history of activity between places at a given moment.

Tableau Machine (p. 27): an interactive Art installation for the home.

Tracking (p. 129): the user task of translating video into English accounts of activity of one target subject including places, objects, people, and interactions.

Trajectory of Appreciation (p. 57): timeline of the degree to which householders engage and use a technology device over a long period.

Translation (p. 40): motion that changes place.

Time to Task Completion (p. 131): the period between the start and end of a task, including its subtasks, and the self-evaluation of results until the operator is satisfied.

Vibration (p. 40): motion that does not change place.

Video Cube (p. 111): a three-dimensional view of video frame sequences mapping time to height.

Video Player (p. 109): a standar video playback console.

Viz-A-Vis (p. 84): interactive visualization of activity through computer vision.

REFERENCES

Abowd, G., G. McGee, M. Morrier, M. Romero, P. Wang, N. Anwer and Y. Hang (2009). Technology and Autism Georgia Tech and Marcus Institute Retreat. Atlanta.

Adelson, E. H. and J. R. Bergen (1985). "Spatio-temporal Energy Models for the Perception of Motion." Journal of the Optical Society of America: 284-299.

Aipperspach, R., E. Cohen and J. Canny (2006). Modeling Human Behavior from Simple Sensors in the Home. Pervasive Computing, Dublin, Ireland, Springer.

Allen, J. E., C. I. Guinn and E. Horvtz (1999). "Mixed-Initiative Interaction." Intelligent Systems and Their Applications, IEEE **14**(5): 14-23.

Amar, B., J. Eagan and J. Stasko (2005). Low-Level Components of Analytic Activity in Information Visualization. IEEE Symposium on Information Visualization, 2004. INFOVIS 2005, Minneapolis, MN, IEEE.

Amar, R. and J. Stasko (2004). A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations. IEEE Symposium on Information Visualization, 2004. INFOVIS 2004. Austin, TX, IEEE.

Aoki, P. M. and A. Woodruff (2005). Making Space for Stories: Ambiguity in the Design of Personal Communication Systems. Proceedings of the SIGCHI Conference on Human factors in Computing Systems. Portland, Oregon, USA, ACM Press.

Barley, S. R. (1990). "Images of Imaging: Notes on Doing Longitudinal Field Work." Organization Science. Special Issue: Longitudinal Field Research Methods for Studying Processes of Organizational Change **1**(3): 220-247.

Barnes, R. (1980). Motion and Time Study: Design and Measurement of Work, Wiley.

Bechtel, R. B. (1997). Environment and Behavior: An Introduction, Sage Publications, Inc.

Bell, G., M. Blythe and P. Sengers (2005). "Making by Making Strange: Defamiliarization and the Design of Domestic Technologies." ACM Trans. Comput.-Hum. Interact. **12**(2): 149-173.

Bennett, E. P. and L. McMillan (2003). Proscenium: a Framework for Spatio-Temporal Video Editing. Proceedings of the eleventh ACM International Conference on Multimedia. Berkeley, CA, USA, ACM.

Bentley, F., K. Tollmar, D. Demirdjian, K. Koile and T. Darrell (2003). "Perceptive Presence." IEEE Comput. Graph. Appl. **23**(5): 26-36.

Blackwell, A. F. (2006). "The Reification of Metaphor as a Design Tool." ACM Trans. Comput.-Hum. Interact. **13**(4): 490-530.

Blythe, M. A., K. Overbeeke, A. Monk and P. Wright (2003). Funology : from Usability to Enjoyment. Boston Kluwer Academic Publishers.

Bobick, A. (1997). "Movement, Activity and Action: the Role of Knowledge in the Perception of Motion." Royal Society Workshop on Knowledge-based Vision in Man and Machine **352**: 1257-1265.

Bobick, A. and J. Davis (1996). Real-time Recognition of Activity using Temporal Templates. 3rd IEEE Workshop on Applications of Computer Vision (WACV '96).

Boehlen, M. and M. Mateas (1998). "Office Plant# 1: Intimate Space and Contemplative Entertainment." Leonardo: Journal of the International Society for Arts, Sciences, and Technology **31** (5): 345-348.

Bolter, J. and D. Gromala (2003). Windows and Mirrors: Interaction Design, Digital Art, and the Myth of Transparency. Cambridge, The MIT Press.

Button, G. and P. Dourish (1996). Technomethodology: Paradoxes and Possibilities. CHI '96: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Vancouver, British Columbia, Canada, ACM**:** 19-26.

Card, S., J. Mackinlay and B. Shneiderman (1999). Readings in Information Visualization: Using Vision to Think. San Francisco, Calif., Morgan Kaufmann Publishers.

Card, S. K., T. P. Moran and A. Newell (1983). The Psychology of Human-Computer Interaction. Hillsdale, N.J., L. Erlbaum Associates.

Cassinelli, A. (2005). Khronos Projector. Los Angeles, ACM SIGGRAPH Emergin Technologies.

Champandard, A. J. (2003). AI Game Development: Synthetic Creatures with Learning and Reactive Behaviors, New Riders Games.

Chen, C. and Y. Yu (2000). "Empirical Studies of Information Visualization: a Meta-Analysis." INTERNATIONAL JOURNAL OF HUMAN COMPUTER STUDIES **53**(5): 851-866.

Coyne, C. (2005). Context Free Art.

Crabtree, A., T. Hemmings and T. Rodden (2002). Pattern-Based Support for Interactive Design in Domestic Settings. Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. London, England, ACM Press.

Daniel, G. and M. Chen (2003). Video Visualization. Proceedings of the 14th IEEE Visualization 2003 (VIS'03), IEEE Computer Society.

Davidhazy, A. (1976). TIME AND SPACE, a one man show held at the Dutchess Community College, Hudson Gallery Poughkeepsie, New York

Dourish, P. (2001). Where the Action Is: The Foundations Of Embodied Interaction. Cambridge, The MIT Press.

Eco, U. (1979). A Theory of Semiotics, Indiana University Press.

Fayyad, U., G. Grinstein and A. Wierse (2002). Information Visualization in Data Mining and Knowledge Discovery. San Francisco, CA, Morgan Kaufmann.

Fels, S., E. Lee and E. Mase (2000). Techniques for Interactive Video Cubism. Proceedings of ACM Multimedia.

Fitzpatrick, G., W. J. Tolone and S. M. Kaplan (1995). "Work, Locales and Distributed Social Worlds." Proceedings of the Fourth European Conference on Computer-Supported Cooperative Work: 1-16.

Fleischman, M., P. DeCamp and D. Roy (2006). Mining Temporal Patterns of Movement for Video Content Classification. Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval.

Gaver, W., J. Bowers, A. Boucher, H. Gellerson, S. Pennington, A. Schmidt, A. Steed, N. Villars and B. Walker (2004). The drift table: designing for ludic engagement. CHI.

Gaver, W., P. Sengers, T. Kerridge, J. Kaye and J. Bowers (2007). Enhancing Ubiquitous Computing with User Interpretation: Field Testing the Home Health Horoscope. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. San Jose, California, USA, ACM Press.

Glasser, B. and A. Strauss (1967). The Discovery of Grounded Theory. Chicago, Aldine.

Google. (2009). "Picasa Image Viewer." from http://picasa.google.com/.

Google. (2009). "Sketchup 7.0." from http://sketchup.google.com/.

Grant, L. and A. N. Evans (1994). Principles of Behavior Analysis. New York, HarperCollins College Publishers.

Gunderson, J. and L. Gunderson (2008). Robots, Reasoning, and Reification, Springer Publishing Company.

Hare, J., P. Lewis, P. Enser and C. Sandom (2006). Mind the Gap: Another Look at the Problem of the Semantic Gap in Image Retrieval. Multimedia Content Analysis, Management and Retrieval 2006, San Jose, California.

Hasbro. (2009). "Cranium." from http://www.hasbro.com/games/cranium/home.cfm.

Hayes, G. (2006). Documenting and Understanding Everyday Activities through the Selective Archiving of Live Experiences. CHI '06 extended abstracts on Human Factors in Computing Systems. Montreal, Quebec, Canada, ACM.

Hillier, B. (1996). Space is the Machine. Cambridge Cambridge University Press.

Hilpoltsteiner, M. (2005). Recreating Movement. Communication Arts. Wuerzburg, Germany, University of Applied Sciences Wuerzburg.

Höök, K., P. Sengers and G. Andersson (2003). Sense and Sensibility: Evaluation and Interactive Art. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Ft. Lauderdale, ACM Press.

Huang, J. (2005). Interactive Wallpaper Proceedings of the ACM SIGGRAPH 05 Electronic Art and Animation Catalog Los Angeles, California ACM Press.

Hughes, J. A., J. O'Brien, T. Rodden, M. Rouncefield and S. Viller (2000). "Patterns of Home Life: Informing Design for Domestic Environments." Personal Technologies **4**: 25-38.

Huhtamo, E. (1998). Silicon Remembers Ideology, or David Rokeby's Meta-Interactive Art. Catalog essay for The Giver of Names exhibit at McDonald-Stewart Art Center.

Huynh, T., B. Ulf and B. Schiele (2007). Scalable Recognition of Daily Activities with Wearable Sensors. Location and Context Awareness, Munich, Germany, Springer.

Iachello, G. and G. Abowd (2005). Privacy and Proportionality: Adapting Legal Evaluation Techniques to Inform Design in Ubiquitous Computing. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Portland, Oregon, USA, ACM.

Intille, L. B. S. S. (2004). Activity Recognition from User-Annotated Acceleration Data. Pervasive Computing, Vienna, Austria, Springer.

ISO (2001). ISO 9126: Software Engineering -- Product quality -- Part 1: Quality model.

Ivanov, Y., C. Wren, A. Sorokin and I. Kaur (2007). "Visualizing the History of Living Spaces." IEEE Transactions on Visualization and Computer Graphics **13**(6): 1153-1160.

Jaschko, S. (2003). Space-Time Correlations Focused in Film Objects and Interactive Video. Future Cinema: The Cinematic Imaginary after Film, MIT Press.

Jeremijenko, N. (1995). Live Wire. Palo Alto, CA.

Joo Geok Tan, D. Z., Xiaohang Wang, Heng Seng Cheng (2005). Enhancing Semantic Spaces with Event-Driven Context Interpretation. Pervasive Computing, Munich, Germany, Springer.

Kapler, T. and W. Wright (2004). GeoTime Information Visualization. IEEE Symposium on Information Visualization, 2004. INFOVIS 2004. Austin, Texas.

Kidd, C., R. Orr, A. G., A. C., I. Essa, B. MacIntyre, E. Mynatt, T. Starner and W. Newstetter (1999). The Aware Home: A Living Laboratory for Ubiquitous Computing Research. Proceedings of the Second International Workshop on Cooperative Buildings - CoBuild'99.

Kiranyaz, S., K. Caglar, E. Guldogan, O. Guldogan and M. Gabbouj (2003). MUVIS: a Content-Based Multimedia Indexing and Retrieval Framework. Seventh International Symposium on Signal Processing and its Applications, ISSPA 2003, Paris, France.

Klein, A. W., P.-P. J. Sloan, A. Finkelstein and M. F. Cohen (2002). Stylized Video Cubes. Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer animation. San Antonio, Texas, ACM.

Koile, K., K. Tollmar, D. Demirdjian, H. Shrobe and T. Darrell (2003). "Activity Zones for Context-Aware Computing." Lecture Notes in Computer Science **2864**: 90-106.

Koile, K., K. Tollmar, D. Demirdjian, H. Shrobe and T. Darrell (2003). Activity Zones for Context-Aware Computing UbiComp 2003: Ubiquitous Computing, Seattle, Washington, Springer.

Krause, J. (2002). Color Index: Over 1100 Color Combinations, CMYK and RGB Formulas, for Print and Web Media, How.

Kubat, R., P. DeCamp, B. Roy and D. Roy (2007). TotalRecall: Visualization and Semi-Automatic Annotation of Very Large Audio-Visual Corpora. Ninth International Conference on Multimodal Interfaces (ICMI 2007).

Kwan, M. P. and J. Lee (2004). Geovisualization of Human Activity Patterns using 3D GIS: a Time-Geographic Approach. Spatially Integrated Social Science: Examples in Best Practice. M. F. Goodchild and D. G. Janelle. New York, Oxford University Press**:** 48–66.

Larson, W. (1967). Figure in Motion.

Lioret, A. (2005). Being Paintings Proceedings of the ACM SIGGRAPH 05 electronic art and animation catalog Los Angeles, California, ACM Press.

Lofland, J. and L. Lofland (1995). Analyzing Social Settings: A Guide to Qualitative Observation and Analysis, Wadsworth Publishing Company.

Mangold, S. C. (2009). INTERACT**:** Multimedia Video Analysis for behavioral research.

Manning, C. and H. Schütze (2002). Foundations of Statistical Natural Language Processing, MIT Press.

Mateas, M. (2001). "Expressive AI: A Hybrid Art and Science Practice." Leonardo: Journal of the International Society for Arts, Sciences, and Technology **34** (2): 147-153.

Mateas, M., T. Salvador, J. Scholtz and D. Sorensen (1996). Engineering Ethnography in the Home. Vancouver, British Columbia, Canada, ACM Press.

Mateas, M. and A. Stern (2003). Facade: An Experiment in Building a Fully-Realized Interactive Drama. Game Developer's Conference: Game Design Track, San Jose, California.

Max-Planck, I. f. P. D. (2009). ELAN**:** Text, audio and video analysis.

McCorduck, P. (1991). Aaron's Code: Meta-Art, Artificial Intelligence, and the Work of Harold Cohen. New York, W.H. Freeman and Company.

Microsoft. (2003). "Windows XP." from http://www.microsoft.com.

Mittelstaedt, E. (2002). Unfolding.

Muhr, T. (2009). Atlas TI**:** Text, digital Audio and Video Analysis.

Munguia, E., S. Tapia, S. Intille and K. Larson (2004). Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors, Berlin Heidelberg: Springer-Verlag.

Mynatt, E. D., J. Rowan, S. Craighill and A. Jacobs (2001). Digital Family Portraits: Supporting Peace of Mind for Extended Family Members. Proceedings of the SIGCHI conference on Human Factors in Computing Systems. Seattle, Washington, United States, ACM Press.

Nagel, H. H. (1988). "From Image Sequences towards Conceptual Descriptions." Image and Vision Computing **6**(2): 59-74.

Nardi, B. A. (1996). Context and Consciousness: Activity Theory and Human-Computer Interaction. Cambridge, Mass., MIT Press.

Neustaedter, C., S. Greenberg and M. Boyle (2005). "Blur Filtration Fails to Preserve Privacy for Home-Based Video Conferencing." ACM Trans. Comput.-Hum. Interact. (TOCHI). **13**(1): 1-36.

Noldus, I. T. (2009). EthoVision 2.3**:** Digital Audio and Video Analysis

Northey, N. W. (1916). "The Angle of View of your Lens." The Camera: An Illustrated Magazine Devoted to the Advacement of Photography **20**(9): 473-484.

Pantic, M., A. Pentland, A. Nijholt and T. Huang (2007). Human Computing and Machine Understanding of Human Behavior: A Survey. Artifical Intelligence for Human Computing, Springer Berlin / Heidelberg. **4451/2007:** 47-71.

Patel, S. N., M. S. Reynolds and G. D. Abowd (2008). Detecting Human Movement by Differential Air Pressure Sensing in HVAC System Ductwork: An Exploration in Infrastructure Mediated Sensing. Pervasive 2008.

Penny, S. (1997). Embodied Cultural Agents at the Intersection of Robotics, Cognitive Science and Interactive Art. Working notes of the Socially Intelligent Agents Symposium., Menlo Park, California, AAAI Press.

Philipose, M. F., K.P. Perkowitz, M. Patterson, D.J. Fox, D. Kautz, H. Hahnel, D. (2004). "Inferring Activities from Interactions with Objects." Pervasive Computing **3**(4): 50-57.

Plaisant, C. (2004). The Challenge of Information Visualization Evaluation. Proceedings of the Working Conference on Advanced Visual Interfaces. Gallipoli, Italy, ACM.

Pousman, Z., Romero, M., Smith, A., Mateas, M. (2008). Living with Tableau Machine: a Longitudinal Investigation of a Curious Domestic Intelligence. In Proceedings of the 10th international Conference on Ubiquitous Computing, UbiComp '08. Seoul, Korea.

Prabhakar, S., S. Pankanti and A. K. Jain (2003). "Biometric Recognition: Security and Privacy Concerns." IEEE Security & Privacy **1**(2): 33-42.

Proshansky, H. (1976). Environmental Psychology: People and Their Physical Settings, Holt McDougal.

Quan, Y., Z. Dong-chi, D. Wan-li, L. Bon-nan and Q. Zheng (2005). Design of an Integrative Automotive Ergonomics Experiment Platform. IEEE International Conference on Vehicular Electronics and Safety, 2005.

Rama Chellappa, S. K. Z., Amit K. Roy-chowdhury (2005). Recognition of Humans And Their Activities Using Video, Morgan & Claypool.

Ramos, G. and R. Balakrishnan (2003). Fluid Interaction Techniques for the Control and Annotation of Digital Video. Proceedings of UIST 2003 – the ACM Symposium on User Interface Software and Technology.

Rode, J., E. Toye and A. Blackwell (2004). "The fuzzy Felt Ethnography: Understanding the Programming Patterns of Domestic Appliances." Personal Ubiquitous Comput. **8**(3-4): 161-176.

Romero, M., Z. Pousman and M. Mateas (2007). "Alien Presence in the Home: the Design of Tableau Machine." Personal and Ubiquitous Computing **12**(5).

Romero, M., J. Summet, J. Stasko and G. Abowd (2008). "Viz-A-Vis: Toward Visualizing Video through Computer Vision." IEEE Transactions on Visualization and Computer Graphics **14**: 1261 - 1268.

Roy, D., R. Patel, P. DeCamp, R. Kubat, M. Fleischman, B. Roy, N. Mavridis, S. Tellex, A. Salata, J. Guiness, M. Levit and P. Gorniak (2006). The Human Speechome Project. Twenty-eighth Annual Meeting of the Cognitive Science Society.

Rozin, D. (1999). Wooden Mirror**:** 830 square pieces of wood, 830 servo motors, control electronics, video camera, computer, wood frame.

Saraiya, P., C. North and K. Duca (2004). An Evaluation of Microarray Visualization Tools for Biological Insight. IEEE Symposium on Information Visualization, 2004. INFOVIS 2004. Austin, TX, IEEE.

Sauter, J. and D. Lüsebrink (1995-2007). Invisible Shape of Things Past. Karlsruhe, Germany.

Seale, A. (1995). Temporal Forms.

Seidman, I. (1998). Interviewing as Qualitative Research: A Guide for Researchers in Education And the Social Sciences. New York, Teachers College Press.

Shannon, C. E. (2001). "A Mathematical Theory of Communication." SIGMOBILE Mob. Comput. Commun. Rev. **5**(1): 3-55.

Shih, T. K. (2002). Distributed Multimedia Databases: Techniques and Applications, IGI Global.

Shneiderman, B. and C. Plaisant (2006). Strategies for Evaluating Information Visualization tools: Multi-Dimensional In-Depth Long-Term Case Studies. BELIV '06: Proceedings of the 2006 AVI workshop on Beyond time and errors. Venice, Italy, ACM**:** 1-7.

Skiljan, I. (2009). "IrfanView." from http://www.irfanview.com/.

Snidaro, L., C. Micheloni and C. Chiavedale (2005). "Video Security for Ambient Intelligence." IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans **35**(1): 133-144.

Sonka, M., V. Hlavec and R. Boyle (1999). Image Processing, Analysis, and Machine Vision, PWS Publishing.

Stiny, G. (1972). Shape Grammars and the Generative Specification of Painting and Sculpture. Proceedings of IFIP Congress.

Stiny, G. (2008). Shape: Talking about Seeing and Doing.

Sturken, M. and L. Cartwright (2001). Practices of Looking: an Introduction to Visual Culture. Oxford, Oxford University Press.

Summet, J., M. Flagg, T. Cham, J. Rehg and R. Sukthankar (2007). "Shadow Elimination and Blinding Light Suppression for Interactive Projected Displays." IEEE Transactions on Visualization & Computer Graphics (TVCG) **13**: 508-17.

Sung, M., C. Marci and A. Pentland (2005). "Wearable Feedback Systems for Rehabilitation." Journal of NeuroEngineering and Rehabilitation **2**(17): 2-17.

Tan, J. G., D. Zhang, X. Wang and H. S. Cheng (2005). Enhancing Semantic Spaces with Event-Driven Context Interpretation. Pervasive Computing, Munich, Germany, Springer.

Tapia, E. M., S. S. Intille and K. Larson (2004). Activity Recognition in the Home Using Simple and Ubiquitous Sensors. Pervasive Computing, Vienna, Austria, Springer.

Terry, M., G. Brostow, G. Ou, J. Tyman and D. Gromala (2004). Making space for time in time-lapse photography. ACM SIGGRAPH 2004 Sketches. Los Angeles, California, ACM.

Terry, M., G. J. Brostow, G. Ou, J. Tyman and D. Gromala (2004). Making Space for Time in Time-Lapse Photography. ACM SIGGRAPH 2004 Sketches. Los Angeles, California, ACM.

Tian, Y.-l., A. Hampapur, L. Brown, L. Feris, M. Lu, A. Senior, C.-f. Shu and Y. Zhai (2009). Event Detection, Query, and Retrieval for Video Surveillance. Artificial Intelligence for Maximizing Content Based Image Retrieval. Z. Ma, IGI Global.

Tinapple, D. (2002). Volumetric Photography.

Truong, B. T. and S. Venkatesh (2007). "Video Abstraction: A Systematic Review and Classification." ACM Trans. Multimedia Comput. Commun. Appl. **3**(1): 3.

Truong, K., G. Abowd and J. Brotherton (2001). Who, What, When, Where, How: Design Issues of Capture and Access Applications. Proceedings of the 3rd

international conference on Ubiquitous Computing. Atlanta, Georgia, USA, Springer-Verlag.

Underhill, P. (2000). Why We Buy: The Science Of Shopping, Simon & Schuster.

Utterback, C. (2004). Untitled 5.

Utterback, C. (2005). Text Rain. Proceedings of the ACM SIGGRAPH 05 electronic art and animation catalog. Los Angeles, California, ACM Press.

Valiant, L. (1984). "A Theory of the Learnable." Communications of the ACM **27**(11): 1134-1142.

Van der Voordt, D. J. M. and H. B. R. Wegen (2005). Architecture in Use: an Introduction to the Programming, Design and Evaluation of Buildings, Architectural Press.

Viegas, F. B., E. Perry, E. Howe and J. Donath (2004). "Artifacts of the Presence Era: Using Information Visualization to Create an Evocative Souvenir." IEEE Information Visualization: 105 - 111

Whyte, W. (1980). The Social Life of Small Urban Spaces, Project for Public Spaces Inc.

William, G. G., B. Robert, W. B. Steven and T. Tan Minh (2003). A Component Architecture for an Extensible, Highly Integrated Context-Aware Computing Infrastructure. Proceedings of the 25th International Conference on Software Engineering. Portland, Oregon, IEEE Computer Society.

Wyche, S., A. Taylor and J. Kaye (2007). Pottering: a Design-Oriented Investigation. CHI '07 Extended Abstracts on Human Factors in Computing Systems. San Jose, CA, USA, ACM.

Zimmerman, A. and M. Martin (2001). "Post-Occupancy Evaluation: Benefits and Barriers." Building Research & Information **29**(2): 168-174.